

# Dynamic evolution of precise regulatory encodings creates the clustered signature of developmental enhancers

Albert Erives<sup>1\*</sup> and Justin Crocker<sup>1</sup>

<sup>1</sup> Dept. of Biological Sciences, Dartmouth College, Hanover, NH, U.S.A.

\* E-mail: Albert.J.Erives@Dartmouth.edu

(Dated: April 7th, 2010)

A morphogenic protein known as Dorsal patterns the embryonic dorsoventral body axis of *Drosophila* by binding to transcriptional enhancers across the genome. Each such enhancer activates a neighboring gene at a unique threshold concentration of Dorsal. The presence of Dorsal binding site clusters in these enhancers and of similar clusters in other enhancers has motivated models of threshold-encoding in site density. However, we found that the precise length of a spacer separating a pair of specialized Dorsal and Twist binding sites determines the threshold-response. Despite this result, the functional range determined by this spacer element as well as the role and origin of its surrounding Dorsal site cluster remained completely unknown. Here, we experiment with enhancers from diverse *Drosophila* genomes, including the large uncompact genomes from *ananassae* and *willistoni*, and report three major interdependent results. First, we map the functional range of the threshold-encoding spacer variable. Second, we show that the majority of sites at the cluster are non-functional divergent elements that have been separated beyond the encoding's functional range. Third, we verify an evolutionary model involving the frequent replacement of a threshold encoding, whose precision is easily outdated by shifting accuracy. The process by which encodings are replaced by newer ones is facilitated by the palindromic nature of the Dorsal and Twist binding motifs and by intrinsic repeat-instability in the specialized Twist binding site, which critically impacts the length of the spacer linking it to Dorsal. Over time, the dynamic process of selective deprecation and replacement of encodings adds to a growing cluster of deadened elements, or *necro-elements*, and strongly biases local sequence composition. Necro-element plaques are associated with mature enhancers that are older than 10 My but not with newer lineage-specific enhancers that employ identical logic. We conclude that the clustered signature of most enhancers results from long histories of selective "maintenance" of precise encodings via facile deprecation and equally facile replacement.

## Introduction

Nothing Gold Can Stay

Nature's first green is gold,  
Her hardest hue to hold.  
Her early leaf's a flower;  
But only so an hour.  
Then leaf subsides to leaf.  
So Eden sank to grief,  
So dawn goes down to day,  
Nothing gold can stay.

—Robert Frost, *New Hampshire* (1923)

How genetic information is encoded in DNA is a central question in biology. In many cases, natural selection acts efficaciously on regulatory DNA sequences, which specify the precise conditions under

which a gene product is made by a cell [1–11]. However, unlike the precise protein-encoding scheme, few general principles have emerged for regulatory encoding. The identification of such principles would facilitate understanding of genomic regulatory DNAs and advance many areas of biological investigation.

One general feature of regulatory DNAs, which include the transcriptional *enhancers*, is the use of combinatorial codes of transcription factor (TF) binding sites [12]. This feature allows an enhancer to activate its gene only if it binds a specific combination of different TF proteins. A less understood general feature is the clustering of multiple binding sites for a single TF operating at an enhancer [13]. This unexplained cluster signature has motivated several bioinformatic screens that exploit binding site density to identify functional enhancers [14,15]. Such

methods detect both functional enhancers and non-functional sequences. Moreover, these methods are not yet predictive of the exact responses encoded by active enhancers bearing site clusters.

Concentration-specific threshold responses are a property of most regulatory DNAs that function through recruitment of DNA-binding factors [16]. However, developmental enhancers that read classical morphogen concentration gradients [17] are ideal subjects in decoding regulatory DNA sequences, and their functional features. Different enhancers with variably-dense clusters of binding sites for the same TF are each responsive to their own unique threshold concentration. Such DNAs can be studied comparatively to identify the variables that encode the concentration threshold setting. In principle, such a variable might be encoded in one of several non-exclusive categories: i) the formulaic combination of adjacent binding sites for TFs acting synergistically; ii) the range of sequences that determine the affinity or allostery of a DNA-bound TF (*functional grammars*); and iii) the higher-order organizational arrangement of binding sites (*functional syntaxes*).

Two well-studied systems of morphogen-responsive enhancers are those that read the Bicoid and Dorsal morphogen concentration gradients that pattern the anterior/posterior (A/P) and dorsal/ventral (D/V) axes of the *Drosophila* embryo, respectively [18–29]. Like most enhancers, these DNAs contain clusters of binding sites, which in this case correspond to those for Bicoid, Dorsal, and their DNA-binding co-factors. This clustering has prompted several complex “cluster code” models that integrate site number, quality, and density parameters to determine the threshold read-out [30–32]. Paradoxically however, the apparent phenotypic robustness of this “cluster code” to mutational divergence has been taken to mean that this full parameter set is simultaneously flexible and determinative [33–36].

To address how concentration-threshold responses are encoded in Dorsal target enhancers, we asked whether there exist a unique, subset of *specialized* TF binding sites in co-clusters of sites for Dorsal, Twist, and Suppressor of Hairless [Su(H)] [37]. Specialized binding motifs, as identified across equivalent enhancers present in a genome and across related lineages, do not manifest the full-range of sequences known to be bound by these factors and may signify regulatory sub-functionalizations. With this approach, we identified two different specialized binding sites for Dorsal, as well as specialized binding sites for Twist and Su(H) [37]. Since then, we

have formally referred to DNA sequences that both drive expression in the lateral embryonic ectoderm and contain this particular collection of specialized binding sites as *Neurogenic Ectodermal Enhancers*, or NEEs [7, 37, 38].

We found that the NEE at the *vnd* locus, or NEE<sub>vnd</sub>, is conserved in *Drosophila* and mosquitos. As such it was present in the latest common ancestor of dipterans ~240–270 million years ago (Mya) [39, 40], or at least >200 Mya [41]. We found that conserved “canonical” NEEs occur at the *rho*, *vnd*, *brk*, and *vn* loci across the *Drosophila* genus [7]. As such, the canonical NEEs were acquired prior to *Drosophila* diversification over 40 Mya [7]. We also found a more recently evolved member of this enhancer class, NEE<sub>sog</sub>, in the *sog* locus of the melanogaster subgroup, which began diverging ~20 Mya [7]. Thus, NEE-type regulatory sequences have been evolving at various unrelated loci within a period spanning the last ~250 My.

NEEs function by recruiting both Dorsal, a *rel*-homology domain (RHD)-containing TF, and its synergistic bHLH co-activator Twist, whose expression mirrors the Dorsal morphogen gradient [19, 42–46]. In addition to having sites for Dorsal and Twist, NEEs possess sites for Su(H) and Snail. Su(H) is a highly-conserved TF that mediates transcriptional responses to Notch/Delta signaling [47–50], while Snail is a highly-conserved C<sub>2</sub>H<sub>2</sub> zinc-finger TF that represses activation in the mesoderm [51, 52]. In *D. melanogaster* and closely related species, NEEs also have a binding site for Dip-3 (Dorsal interacting protein-3) [37], a Dorsal-binding protein required for Dorsal/Twist synergistic activation and D/V patterning [53–56]. Besides these specialized binding sites, NEEs share distinct organizational features pertaining to site placement, spacing, and polarity [37]. These observations suggest that NEEs form a distinct set of sequences that “read-out” the Dorsal morphogen gradient at various thresholds in the lateral regions of the embryo through specific protein complexes composed of Dorsal, Twist, Snail, Su(H) and their co-factors.

Recently, we determined that the specialized NEE-type binding sites for Dorsal and Twist have a unique function in setting the threshold for activation [7]. In the NEEs from *D. melanogaster*, *D. pseudoobscura*, and *D. virilis*, we found that: i) the precise length of a spacer DNA, which separates these well-defined Dorsal and Twist binding sites, encodes the concentration threshold setting; ii) natural selection has acted on the length of this spacer in different lineages of the *Drosophila* genus to adjust the

threshold; and iii) these selective *cis*-regulatory adjustments have been performed at all NEEs across a given genome, as would be expected if they are all co-evolving to a common change in the *trans*-morphogen gradient [7]. While this study identified a heritable feature that encodes different responses to Dorsal, it did not address its full functional range nor the function of the many other Dorsal binding site variants, which constitute the clusters observed at these enhancers. As such, it was not clear whether these additional Dorsal motifs were necessary and/or sufficient for setting the gradient threshold, participating in activation or repression, or any other regulatory function.

Here we test several wild-type and experimentally-modified NEEs from five divergent species of *Drosophila*: *D. melanogaster*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis*. Importantly, *D. ananassae* and *D. willistoni* represent the largest assembled *Drosophila* genomes and are less derived than the smaller, compact genomes of the melanogaster subgroup, which may have lost important signatures indicative of past evolutionary history [57]. Using this broad data set, we narrow the many explanations of binding site clustering down to a single, unexpected, but ultimately predictive hypothesis of concentration-threshold encoding, and explain several perplexing constraints on the specialized sites of NEEs and their relative organization. We show that complex enhancer clustering is a signature that ages over time through a dynamic evolutionary process involving facile selection for optimal threshold readouts and equally facile loss and/or selective deprecation of former threshold-encodings. This process, which we term *dynamic deprecation*, produces several non-functional signatures that obscure the precise morphogen threshold-encoding mechanism that we functionally map and confirm in this study. We conclude that the clustered signature observed in most enhancers is produced by the dynamic evolutionary maintenance of the accuracy of precise threshold-encodings.

## Results

### Canonical NEEs are marked by *cis*-spectral clusters

We found that binding site clusters at NEEs are characterized by a certain “*cis*-spectral” signature, and refer to such clusters simply as *cis*-spectra (Fig. 1). Binding site constituents of *cis*-spectra are revealed specifically within or immediately around

the cluster as the motif consensus for a TF is relaxed. Thus, a *cis*-spectral binding cluster remains well-defined with increasing degeneracy of the binding motif. For example, if we use a *motif spectrum* of increasingly degenerate binding motifs characteristic of Dorsal binding sites, we identify additional matching sequences locally within the vicinity of the module, thus preserving the definition of the cluster (Fig. 1, bottom rows of localized clustering).

We defined three specialized *cis*-element motifs that are associated with the *cis*-spectral clusters of canonical NEEs across *Drosophila*: *SUH/D $\alpha$* , *D $\beta$* , and *E(CA)T* (see Fig. 1). These motif signatures correspond to specialized versions of more general binding motifs for Dorsal, Twist, Snail, and Su(H). Importantly, the specialized motifs typically describe a single site at each *cis*-spectral cluster.

Despite the numerous binding site variants in Dorsal *cis*-spectra, there are only two distinct and separate, specialized Dorsal binding site motifs at each NEE, here called *D $\alpha$*  and *D $\beta$* . The specialized Dorsal binding motif *D $\alpha$*  partially overlaps an overly-determined and polarized Su(H) binding site *SUH* (Fig. 1). In *D. melanogaster*, *SUH* is polarized in the same direction as the  $\mu$  site, a specialized binding site for Dip-3 [37]. Furthermore, while the  $\mu$  element appears to be absent in distant *Drosophila* lineages, *SUH* is maintained in a polarized state, even after turnover events [7].

In contrast, the specialized Dorsal binding motif *D $\beta$*  is located uniquely within  $\sim 20$  bp of the *E(CA)T* element, the spacing to which encodes the threshold response. Furthermore, an invariant length-asymmetry in this nearly palindromic *E(CA)T* motif consistently points to *D $\beta$*  although *D $\beta$*  itself is not polarized. Importantly, we have never observed any Dorsal binding site variant to be more tightly linked to the *E(CA)T* element than the *D $\beta$*  element. The *E(CA)T* element itself is a specialized CA-core E-box (5'-CANNTG) with an additional T, i.e. the sequence 5'-CACATGT. This *E(CA)T* element is partially explained as the superimposition of binding preferences for Twist and Snail. Activating Twist:Daughterless bHLH heterodimers bind the YA-core E-box 5'-CAYATG, or *E(YA)*, while the Snail repressor binds the motif 5'-SMMCWTGYBK [51, 58]. Thus, we predicted that such a co-functional site may originate via selection for the superimposed motifs, which corresponds to the sequence 5'-SCACATGY. This superimposed Twist/Snail binding motif is almost identical with the observed *E(CA)T* motif, 5'-CACATGT.

We will refer to the three arranged elements of

the polarized  $E(CA)T$  site, the spacer, and an unpolarized  $D\beta$  site as an  $E$ -to- $D$  encoding. Using this terminology, we will show that functional NEE modules need be composed only of one  $E$ -to- $D$  encoding, supported by a nearby generic Su(H) site. We will show that the  $E$ -to- $D$  sequence is the sole repository of the threshold encoding variable at each NEE module, and that *cis*-spectral clusters and certain specialized sites are byproducts accumulated in mature enhancers. Last we will show that an intrinsic mutational property of the  $E(CA)T$  elements facilitates the rapid selection of new  $E$ -to- $D$  encodings.

### Canonical NEEs from *D. willistoni* genome are enriched in *cis*-spectra

To better understand the functional importance of multiple variant binding sites for Dorsal and its co-factors within canonical NEEs, we analyzed the *D. willistoni* genome, which is the largest assembled *Drosophila* genome (224 Mb) [57]. The study of large genomes is important because relatively compact genomes may have lost DNA signatures indicative of past evolutionary processes. The *D. willistoni* lineage is an early branch of the same SOPHOPHORA subgenus that includes the *melanogaster* subgroup, and represents  $\sim 37$  My of evolutionary divergence since its common ancestor with *D. melanogaster*, whose genome has been secondarily compacted (Fig. 2).

To identify the canonical NEE set from *D. willistoni*, it is sufficient to query the genome for all 800 bp sequences containing the three motifs given by  $SUH/D\alpha$ ,  $D\beta$ , and  $E(CA)T$ , without imposing any syntactical constraints, such as linked Dorsal/Twist binding sites or polarized  $SUH$  elements. Such a query identifies only the four canonical NEEs of *Drosophila*, and these all conform to the full syntactical rule set, despite significant levels of sequence divergence. We also verified that these NEE-bearing loci are expressed in the neurogenic ectoderm of *D. willistoni* embryos by whole-mount *in situ* hybridization (Fig. 3 A–D).

We cloned DNAs encompassing the NEE sequences of *D. willistoni* and tested them for enhancer activity on a *lacZ* reporter stably integrated into multiple independent lines of *D. melanogaster*. Whole-mount *in situ* hybridization of these embryos with an anti-sense *lacZ* probe showed that the *D. willistoni* enhancers drive lateral ectodermal expression in *D. melanogaster* embryos (Fig. 3 E–H). These results demonstrate that these are functional enhancers present in loci expressed in lateral regions

of the neuroectoderm in *D. willistoni* embryos. In general, *D. willistoni* NEEs drive slightly narrower expression patterns in *D. melanogaster* than their counterpart *D. melanogaster* NEE reporters, which may indicate that they are tuned to higher threshold responses (Fig. 4).

To determine whether the specialized Dorsal binding sites  $D\alpha$  and  $D\beta$  are embedded in clusters of Dorsal binding site variants as they are in other lineages, we identified all sites in these sequences matching a Dorsal motif spectrum and found extremely dense Dorsal *cis*-spectra in the NEEs of *D. willistoni* (Figs. 5–6). As quantified below, these are some of the densest clusters yet seen in NEEs of the *Drosophila* genus. To ascertain whether the specialized Dorsal motifs are maintained as unique copies in each NEE from *D. willistoni*, or whether additional Dorsal binding variant sites within each cluster also match these specialized motifs as would be expected by random neutral drift [59,60], we applied our path-finding method to identify and characterize the most specialized Dorsal binding motifs within their *cis*-spectral clusters [37] (also see Supplement Part I). We find that the  $D\alpha$  site occurs once in each NEE (Fig. 5). Furthermore,  $D\alpha$  continues to overlap the Su(H) binding site at this particular specialized Dorsal binding site. This property is unique to  $D\alpha$  in canonical NEEs across the *Drosophila* genus. Similarly, the  $D\beta$  of *D. willistoni* site occurs only once in each NEE (Fig. 6). As expected,  $D\beta$  is the closest variant Dorsal binding site adjacent to  $E(CA)T$  (Fig. 6). The  $D\beta$  consensus motif for the canonical NEEs of *D. willistoni* is nearly identical with the corresponding motif in other previously-characterized lineages (Table 1).

The Dorsal *cis*-spectral clusters of NEEs from *D. willistoni* are associated with another feature that is interesting in light of the reduced genomic deletion rates relative to *D. melanogaster*: the *D. willistoni* NEEs appear to be enriched in CA-satellite sequence. Given that the  $E(CA)T$  sequence, 5'-CACATGT, is composed entirely of CA-dinucleotide repeats, we speculated whether the Dorsal *cis*-spectra of NEEs are overlaid with a similar  $E(CA)T$  spectral cluster. In support of this idea, we found several lengthy CA-satellite tracts across the canonical NEE set of *D. willistoni* (Fig. 7). Almost all of these are associated with specific constituents of Dorsal *cis*-spectra. Conversely, almost all constituent sites of Dorsal spectra are associated with prominent CA-satellite tracts. For example, the *cis*-spectral cluster of the NEE<sub>vn</sub> of *D. willistoni* has an expanded CA-satellite tracts associated with divergent  $D\beta$  ele-



ments at  $\sim 340$ – $400$  bp and again at  $\sim 580$ – $630$  bp, while the *D. willistoni*  $NEE_{rho}$  also has an expanded CA-satellite tracts coordinated to divergent  $D\beta$  elements at  $\sim 130$ – $150$  bp and again at  $\sim 270$ – $290$  bp (Fig. 7). Last, the  $NEE_{vnd}$  sequence, which is the descendant of the oldest known NEE because it is found in mosquitos, is characterized by the greatest number of lengthy CA-satellite tracts in *D. willistoni* (Fig. 7).

### Constituents of *cis*-spectra represent non-functional necro-elements

In the  $NEE_{vnd}$  module of *D. willistoni*, we detected the loss of one of two *E-to-D* encodings that are present and intact in the  $NEE_{vnd}$  sequences from the *D. melanogaster*, *D. pseudoobscura*, and *D. virilis* genomes [7, 37]. The first *E-to-D* encoding has a tighter spacer compared to the second, distantly-spaced *E-to-D* encoding. Furthermore, the Dorsal binding site at this second encoding is a divergent  $D\beta$  element (Fig. 1). In the *D. willistoni* lineage, the  $E(CA)T$  element of this second divergent encoding expanded on both sides and then split apart (Fig. 8A, inverted CA-satellite palindromic pair #4). This is unambiguously an inactivating mutation of the Twist binding element. Furthermore, the  $NEE_{vnd}$  of *D. willistoni* is marked by several other such palindromic tracts (numbered in Fig. 8A), of which the intact but also expanded  $E(CA)T$  site is the leftmost site in a series of increasingly-lengthy, split, inverted palindromic CA-satellite repeats (Fig. 8B). These increasingly expanded CA-satellite palindromes are associated with Dorsal binding site variants that are increasingly divergent from the  $D\beta$  consensus motif (Fig. 8C).

While the *D. willistoni*  $NEE_{vnd}$  sequence has lost the second  $E(CA)T$  site through repeat expansion and separation of the two palindromic moieties, we did not know whether this site functioned in species in which it is still intact. We therefore tested two different fragments contained within a “full-length” 949 bp enhancer sequence from the *vnd* locus of *D. melanogaster* (Fig. 9A). We tested a 300 bp fragment that contains the first *E-to-D* encoding spaced by 10 bp, and a 266 bp fragment that contains the second *E-to-D* encoding spaced by 20 bp. Both fragments overlap and contain in common the extended *SUH/D $\alpha$*  site (Fig. 9A).

We found that the 300 bp fragment works just as well as the 947 bp fragment (Fig. 9 B, C, and E) while the 266 bp fragment hardly works at all (Fig. 9D and 4E). Thus, the first *E-to-D* encoding,

which is intact and tightly spaced, is sufficient for the complete threshold-response, while the second *E-to-D* encoding, which is expansively-spaced to a slightly divergent  $D\beta$  element, is non-functional. We refer to the component sites of the second encoding as dead elements, or *necro-elements*, and label them  $N-E(CA)T$  and  $N-D\beta$ . While the  $N-E(CA)T$  sequence is intact, inspection of this  $N-D\beta$  sequence shows that it has diverged somewhat from the genus-wide  $D\beta$  consensus (Fig. 9F).

These results indicated that Dorsal *cis*-spectra and their associated CA-satellite tracts are relic *E-to-D* encodings that were once functional but eventually deprecated and replaced during lineage evolution. While the evolution of new encodings will sometimes occur via selection of spacer length variants defined by existing elements, at other times it will occur via selection of new replacement sites associated with new spacer lengths. Three important features of *E-to-D* encodings increase the capacity for selection of replacement encodings. The first feature is the palindromic nature of the  $E(CA)T$  and  $D\beta$  elements, which allows new *E-to-D* encodings to arise from the selection of a single emergent site that is located on the other side of its coordinating partner element in an existing encoding (a leapfrog). The second feature is that the *E-to-D* spacer range is broad-ranged and thus endows functionality to sub-optimal encodings. The third feature is that CA-dinucleotide satellite sequence is susceptible to repeat expansions and contractions across the *Drosophila* genus [61–63]. We assume that the  $E(CA)T$  sequence 5'-CACATGT is dynamically unstable in NEEs because this element is composed entirely of CA-repeats. In support of this, we found that intact  $E(CA)T$  elements in the NEEs of several *Drosophila* genomes are frequently repeat-expanded beyond the core heptamer such that it matches the general pattern given by 5'-(CA) $_n$ T(GT) $_m$ , where  $n \geq 2$  and  $m \geq 1$  (Table 2). This is pronounced particularly in the larger, uncompact *D. ananassae*, *D. willistoni*, and *D. virilis* genomes, (Table 2). These observations are of utmost significance: spacer length variants produced by an intrinsic repeat instability of the  $E(CA)T$  element will drive different threshold-responses. This eventuality would also explain the highly invariant nature of the  $E(CA)T$  sequence. Newly-selected replacement Twist/Snail binding sites will evolve at target sequences most closely resembling the dual-functioning site predicted by superimposed binding preferences (Fig. 10). Initially, such an emergent site will be associated with a suboptimal

spacer. However, random neutral drift to the specific  $E(\text{CA})T$  sequence would result in the availability of spacer length variants via CA-satellite repeat expansion/contraction. Thereafter, frequent occasions for selection of spacer variants produced by such a site would result in the apparent “constraint” of the Twist element.

The evolution of threshold readouts via dynamic deprecation and replacement of encodings, as facilitated by intrinsic  $E(\text{CA})T$  instability, makes several testable predictions. First, a dynamic deprecation model is supported if longer CA-satellite tracts in *D. willistoni* NEEs are loosely associated with specific components of Dorsal *cis*-spectra, especially when they are spaced beyond the functional range of the spacer element. Second, necro-element accumulation may progress in a clock-like fashion followed by neutral divergence of these sites. Thus *cis*-element spectra for both Dorsal and Twist binding motifs should be associated with mature NEEs that are canonical to the lineage, but not in newer NEEs that might have arisen more recently. Third, we should find that threshold readout is correlated to spacer length but not to binding site density. Fourth, we should be able to remove deprecated encodings without affecting the threshold readout (as in Fig. 9 C and E). Conversely isolated deprecated encodings should not possess lower thresholds compared to the intact enhancer (as in Fig. 9 D and E).

### Canonical NEEs across *Drosophila* are enriched in necro-element spectra

To address the generality of CA-satellite accumulation in NEEs across the genus, we checked the percentage of CA-satellite in NEEs from *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis* relative to their genomic background levels (Table 3). These analyses consistently show that CA-satellite is enriched in NEEs above genomic background rates. Importantly, this elevated level is not due to the presence of intact  $E(\text{CA})T$  motifs, which constitute only a minor fraction of CA-repeat sequence in NEEs (Table 3).

To address the possibility that elevated CA-satellite composition is a feature common to developmental enhancers, we then looked at several canonical enhancers that respond to the Bicoid morphogen gradient, which patterns the anterior/posterior (A/P) axis. We identified the *hunchback* (*hb*) enhancers, the *giant* (*gt*) posterior enhancers, the *Kruppel* (*Kr*) enhancers, and the well-studied *even-skipped* (*eve*) stripe 2 enhancers from

each of 4 genomes: *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis*. All of these enhancers are active in the same embryonic nuclei as the NEEs and thus constitute a well-matched control group. We found that while all 16 of these A/P enhancers possess evolving clusters of Bicoid binding site spectra (data not shown), none of them possess the elevated CA-satellite levels that characterize canonical NEEs from these same species (Fig. 11). Thus, there is a tremendous sequence bias that is unique to canonical NEEs across the genus and in stark contrast to the sequence composition of both their genomes and other non-NEE enhancer clusters. Furthermore, this NEE compositional bias is related to specific functional elements employed by NEEs.

Having found we could identify the extent of Dorsal *cis*-spectra with confidence, we then checked its potential to encode or influence Dorsal concentration threshold read-out of NEEs. For example, we checked the relation between threshold-readout and the density of Dorsal halfsites in a region anchored  $\pm 480$  bp from *D $\beta$*  (Fig. 12A). For this we measured the stripe width at 50% egg length as measured by the number of nuclei expressing the reporter gene from the ventral border of expression up to the dorsal border. We also found no relation between Dorsal binding site densities and threshold-encodings after trying diverse other descriptors of a Dorsal binding site (data not shown). Identical densities of Dorsal halfsites, degenerate full-sites, and more complete full-sites are present in different enhancers that read-out different Dorsal concentration thresholds and vice versa.

In contrast, if we plot the length of the *E-to-D* spacers for NEEs with unambiguous *E-to-D* encodings (i.e., encodings with single intact  $E(\text{CA})T$  and *D $\beta$*  elements) and except those from the dorsally-repressed *vnd* loci, we see a well-defined, hump-shaped curve, whose peak activity tops at around 7 bp and falls on either side of this maximum. The spacer elements from the consistently high-threshold *NEE<sub>vnd</sub>* sequences across the genus obey a similar, albeit depressed, curve because of one additional regulatory input (data not shown). Thus, the elevated CA-satellite content and its associated Dorsal *cis*-spectra are consistent with the central hypothesis that the sequence composition of these enhancers has been shaped by a long history of repeated deprecation and compensatory selection of *E-to-D* encodings by a process which has been active for more than 200 My in the case of the *NEE<sub>vnd</sub>* sequence, and more than 40 My at other canonical NEEs.

Given the extent of *cis*-spectral signatures as-

sociated with Dorsal and Twist binding elements in mature NEEs, we asked whether the specialized  $D\alpha$  site, which overlaps an unusually specialized Su(H) binding site, might also be a  $D\beta$  necro-element that was conveniently turned into a Su(H) site. To address this question, we first compared the  $D\alpha$  and  $D\beta$  consensi motifs across all five divergent *Drosophila* lineages for which we functionally tested NEEs in *D. melanogaster* (Table 1, Fig. 13A). Remarkably, we find that the second half of the  $D\alpha$  has diverged across the genus faster than the first half. This second half is the portion that does not overlap the Su(H) binding site. Unlike, the slight lineage-specific variations of  $D\beta$ ,  $D\alpha$  motif divergence can be characterized as increasingly degenerate when departing from the ancestral  $D\alpha$  motif, which is closest to a  $D\beta$  motif itself.

To test whether the Su(H) binding site is itself functional and perhaps the principal reason for persistence of the “ghost”  $D\alpha$  motif, we specifically mutated the Su(H)-specific portion of the *SUH*/ $D\alpha$  site in the  $NEE_{rho}$  of *D. melanogaster* (Fig. 13 A and C). This specific mutation appears to weaken the activation response of the enhancer without affecting the specific threshold setting (Fig. 13 B–C). Because we have shown a general tendency of functional  $E(CA)T$  elements to have expanded beyond the heptamer sequence (Table 2), and of deprecated  $E(CA)T$  elements to have experienced runaway expansion into longer tracts (Figs. 7–8), we suspect that this process tends to push away combinatorial enhancer elements, such as Su(H) binding sites. In this context, selection may favor new Su(H) binding sites that are closer to the current functional encoding. Conveniently, deprecated  $N$ - $D\beta$  sequences are similar to sequences matching the Su(H) binding motif and thus provide a convenient set of target sites for re-evolving more proximal Su(H) sites.

### Newly evolved NEEs are not enriched in *cis*-spectra

Our results on the canonical NEEs of the four divergent lineages of *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis* NEEs demonstrate that much of their sequence composition corresponds to relic deprecated encodings. This pertains not only to the sequences in between intact Dorsal, Twist/Snail, and Su(H) binding motifs but to most of the recognizable and intact TF sites and variants as well. Because we predict that necro-element accumulation is a neutral signature related to the number of past threshold adaptations, whose number likely in-

creases with age, we were curious about the extent of *cis*-spectral signatures in younger NEEs. We previously documented a new NEE sequence at the *sog* locus of *D. melanogaster* [7]. The *D. melanogaster*  $NEE_{sog}$  sequence has a CA-dinucleotide content of 14.4%, which is on par with highest levels seen in A/P enhancers from all lineages but is mid-range for NEEs from *D. melanogaster* (compare with Fig. 11B points in the A/P box). However, because the CA-content of NEEs from *D. melanogaster* may have been secondarily reduced, we therefore wanted to query uncompact *Drosophila* genomes with a parameter set that is constrained only by the minimal molecular requirements. Thus, we queried the two largest *Drosophila* genome assemblies, which corresponded to *D. ananassae* (231.0 Mb) and *D. willistoni* (235.5 Mb). Both of these species are in the SOPHOPHORA subgenus, which includes *D. melanogaster*.

Of the 1 kb genomic windows centered on all  $D\beta$  instances in any given genome and containing  $E(CA)T$  anywhere in that window, we identified the subset of these sequences that also contained a generic (“un-specialized”) Su(H) binding site as well as linked Dorsal and Twist binding elements. The generic Su(H) site replaces the composite extended motif that described an overly-determined *SUH* element and the overlapping  $D\alpha$  ghost site. Using this set of minimal criteria, we nonetheless were able to identify the canonical NEE repertoires for each species.

From the *D. ananassae* genome, we identified, cloned and tested both a functional set of canonical NEEs (Fig. 14), and a new NEE at the *Delta* (Dl) locus (Fig. 15). Delta encodes a ligand for the Notch receptor, whose signaling is relayed by the Su(H) TF itself [49, 64]. In *D. melanogaster* embryos, Delta is expressed in a narrow lateral stripe in the mesectoderm and ventral most row of the neurogenic ectoderm using sequences that are unrelated to the  $NEE_{Delta}$  sequence of *D. ananassae* [50].

Like the  $NEE_{sog}$  sequence, which matured in the melanogaster subgroup, the  $NEE_{Delta}$  sequence in *D. ananassae* has not yet accumulated either CA-satellite content or the Dorsal *cis*-spectra characteristic of necro-element plaques (Fig. 15A). Nonetheless, this enhancer is functional in *D. melanogaster* embryos (Fig. 15B). Inspection of its Su(H) binding site reveals that it does not overlap a ghost  $D\alpha$  motif, which demonstrates again that  $D\alpha$  is not required (Fig. 15C). This is consistent with the interpretation that  $D\alpha$  motifs are deprecated  $D\beta$  motifs exapted into functional *SUH* elements at ma-

ture NEEs, whose sequence compositions have been biased by long histories of necro-element accumulation.

The NEE<sub>Delta</sub> enhancer has a spacer of 3 bp, and occupies the low-end of the threshold mapping function (Fig. 12). Therefore, because we characterized both high and low threshold NEEs that have evolved more recently in the *Delta* and *sog* loci of *D. ananassae* and *D. melanogaster*, respectively, without much necro-element accumulation, the *cis*-spectra of mature NEEs are likely unrelated to function. Instead, the absence of the necro-element plaques suggest a shorter period of evolutionary maintenance, consistent with their phylogenetic distribution.

## Discussion

In this study of regulatory DNAs from the *Drosophila* genus, we found that a certain Dorsal-threshold encoding mechanism maps a spacer length of 3–15 bp, which links a pair of well-defined Dorsal and Twist binding sites, onto one well-defined dorsal border of expression that is 5–15 nuclei past the ventral border of the neurogenic ectoderm. The specialized Twist-binding *E(CA)T* sequence is a constrained motif that satisfies binding preferences for both the Twist activator and the Snail mesodermal repressor. This sequence is also a palindromic CA-satellite sequence that is prone to CA-dinucleotide repeat expansions that alter the precise threshold setting spacer. Natural selection acts continuously to exploit *E(CA)T* instability to adapt the precise, threshold-setting spacers between adjacent and intact Dorsal and Twist binding elements. This process may also accelerate site turnover, because it would frequently necessitate stabilizing selection of compensatory threshold settings in response to this intrinsic instability. Thus, evolutionary maintenance of optimal NEE function involves the clock-like production of dead Dorsal and Twist binding elements, which we call necro-elements. Necro-element accumulation is the major determinant of sequence composition in enhancers that have matured beyond a certain age (>10 My). Further genomic sampling of taxa will allow refinement of the necro-element clock, and ascertain whether it reaches a saturation point for the most ancient enhancers. This question increases the need to sequence larger genomes that are not compressed secondarily by high deletion rates [65].

We found that the specialized Su(H) binding site *SUH* is exapted from deprecated, non-functional Dorsal binding sites in all canonical

NEEs of *Drosophila*. *SUH* appears to influence the strength of activation without affecting the Dorsal concentration-threshold response. This site is specialized in mature NEEs but not in more recently evolved NEEs. This unusual turnover process for Su(H) sites may be necessitated by the tendency of CA-satellite expansion to act as a “conveyor belt” pushing out coordinating elements such as the Su(H) binding site, but leaving a convenient path of deprecated elements that are easily exapted into closer Su(H) sites.

We found that functional NEEs can be derived from truncated fragments of mature NEEs that lack necro-elements while continuing to encode the correct threshold setting. Also, functional NEEs have evolved more recently at non-canonical loci without having yet accumulated the characteristic necro-element plaques seen in older NEEs. Such NEEs bear Su(H) sites that do not extend to deprecated, ghost Dorsal binding sites.

Last, we found a smooth continuum between intact NEE elements and increasingly divergent deprecated necro-elements in these enhancers. Furthermore, because the extreme range of this continuum is associated with the age of the enhancer, we infer that necro-element accumulation begins with each NEE origination and is continuously co-extant with its adaptive maintenance. This has led us to a richly-predictive yet parsimonious model of NEE evolution that we call *dynamic deprecation* (Fig. 16). With increasing time, the background sequence composition of enhancers is profoundly altered and eventually dominates the nature of binding site sequences because it provides a highly-biased ground state from which new sites are exapted.

## Defining necro-elements, *cis*-spectra, and deprecated necro-elements.

We have used the term *necro-element* initially to describe intact or nearly intact binding sites occurring within well-defined clusters but which are no longer relevant in the current threshold encoding. This term can be applied to sites subjected to dynamic deprecation, including those that are deprecated solely through changes in syntax. However, because there is no clear dividing line between potentially-functional binding sites deprecated by syntax and increasingly divergent sites, we have chosen to expand the use of “necro-element” to refer to the entire continuum constituting a clustered plaque of necro-elements. We call such clusters *cis*-spectra in order to distinguish them from functional “clusters” of binding sites. *Cis*-spectra are well-defined operationally as



motif clusters that remain distinct from background genomic sequence as the degeneracy of the matching motif is increased and additional, presumably older, relic sites are revealed. In this context, we used the term *motif spectrum* to refer to the bioinformatic set of motif descriptors that detect *cis*-spectra for a given TF.

The use of the prefix *necro*- rather than the prefix *pseudo*- is justified by several important distinctions that are peculiar to necro-elements. Etymologically, the Greek root  $\psi\epsilon\upsilon\delta\omicron$ - means ‘false’, while the Greek root  $\nu\epsilon\chi\rho\omicron$ - means ‘dead’ and more accurately connotes ‘loss of function’. This is an important distinction because biological systems are rich in functional dissimulation (e.g., mimicry and camouflage on an organismal scale, but also extending to viral oncogenes that dissimulate normal cellular genes, and potentially true pseudo-elements that function as decoy DNA elements to sequester a transcription factor). Biologically, the chosen term must encompass in its definition both deprecated and non-deprecated elements, as well as both non-functional and functionally-redundant elements. Conventionally, the usage of the *pseudo*- prefix for sequence lengths on the length-scale of *cis*-elements is unwieldy because it is used almost exclusively for recognizable homologs of protein-coding genes with clear inactivating mutations (e.g., internal stop codons, and frameshifts). Necro-elements cannot always be identified by sequence alone because they can be rendered functionally redundant or non-functional by selection on syntax.

We also used the term *deprecation* to connote additional information as to the probable role of selection in producing a necro-element. A *deprecated necro-element* is a useful distinction to characterize a necro-element that has undergone selection for attenuated or complete loss of function in connection with the selection of a replacement threshold-encoding located either at the enhancer or elsewhere in the locus. Thus, deprecation implies that selection was active in removing an epistatic relationship between two conflicting threshold-encodings. Selection may favor such an outcome when the pre-deprecated functional element encodes a lower-threshold than the positively selected replacement encoding. In such cases, a low threshold encoding masks the function of any high threshold encoding under positive selection and must engender active selective deprecation. On the other hand, if a high-threshold encoding is being selectively replaced by a low threshold encoding, we expect no active deprecation forces. Instead, we expect gradual loss

of function via neutral drift. This is an unexpectedly novel evolutionary mechanism for generating apparent regulatory redundancy. In this context, we suggest that redundant “shadow enhancers”, which have been observed at several Dorsal target loci in the *D. melanogaster* genome [66], should be incorporated into the same dynamic deprecation framework when appropriate. Selection may adapt an existing threshold encoding or transition its focus to a new threshold encoding that is located either within the same enhancer or elsewhere in the locus. Multiple such events are likely to pepper the idiosyncratic histories of different lineages at different times. In this context, *shadow enhancers* may be defined as out-moded enhancers, which were either redundant when replaced by distant low threshold enhancers, or actively deprecated by selection until their threshold was at least higher than a newer optimal low threshold enhancer located elsewhere in the same locus.

**Summary and implications.** In principle, *cis*-spectral plaques of necro-elements should accumulate in all complex eukaryotic enhancers that encode key regulatory variables in a precise syntax. The extent of this clustering would then be determined by the age of the enhancer, and the number or rate of replacement adaptations over this time. While many of the intensely studied enhancers of *Drosophila* have corresponded to early embryonic enhancers that are evolutionarily sensitive to changes in egg size and morphology, they are also proving useful in untangling the molecular and evolutionary aspects of enhancer biology.

In this evolutionary context, the biology of necro-element spectra of D/V enhancers appears to be directly applicable to A/P enhancers responsive to the Bicoid morphogen gradient system. Evolution of egg size and developmental timing during embryogenesis is likely to place evolutionary demands on both A/P and D/V morphogen gradient systems, which are operating simultaneously in the same cells. While we have shown that Dorsal binding site density does not correlate with threshold encoding, others have shown that Bicoid binding site strength in the heavily-clustered A/P enhancers does not correlate with A/P position of activity [67]. Under the dynamic deprecation theory of enhancer evolution, this paradox is explained if the majority of Bicoid binding site variants at such clusters represent necro-elements deprecated by mutations affecting the site itself, its coordinating site(s), and/or their syntactical relation. This interpretation can be confirmed by future studies identifying the minimal molecular

requirements for encoding variable Bicoid-response thresholds.

One important implication for current studies is that motif descriptors and algorithmic motif predictors should be constructed over a judiciously-chosen set of functionally-equivalent sites across a genome, rather than on the continuum of necro-element spectra at a cluster. Such clusters are often exploited statistically to increase the number of “example elements”. Such approaches lead to degenerate motifs describing both extant functional elements and surrounding deprecated sequences. Newer approaches that are both alignment-free and wary of exploiting the abundance of related sequences will do better at distinguishing functional elements from evolutionary artifact [37, 68].

The conceptual re-framing of the functional evolution of enhancers overturns a common assumption that all binding site variations within an enhancer are functional and/or subtly necessary. This assumption has been directly responsible for the impression that the “cluster code” is “flexible”, by which is meant that enhancer activity is robust to mutational disruption [33–36]. However, whether these site sequences are flexible or not flexible is only a productive question if the observed sequences are functional in some way. In contrast, our results have supported the existence of a precise encoding scheme that uses only a *limited subset* of sites in the cluster [37]. Mutational variation in the organization of these specialized sites produces a specific and well-mapped range of expression phenotypes [7]. Indeed, because this precise encoding scheme turns brittle when extended past its functional range, selective deprecation is facilitated. This view is further enriched by considering the complex macroevolutionary processes that result when taxa and lineages persist through several expansions caused by non-static ecological/climatic conditions [69–72]. Regulatory evolution is likely to underlie many of the stabilizing and adaptive changes associated not only with these climate-driven historical events but future climate changes as well [73].

The potential for gene regulatory evolution is likelier when encoding schemes for relevant regulatory traits are broad-ranged functions that map genotype (enhancer sequence) to phenotype (expression profile). Precise codes provide the additional category of syntax on which natural selection can act. However, a broad or evolutionarily-varied phenotypic range may be a simple consequence of molecular mechanisms that are employed ontogenetically

at multiple loci in precise but varied functional configurations. Understanding this complex relation between molecular encoding systems and their complex evolutionary histories may prove useful in gauging the intrinsic adaptive potential of specific systems subjected to future climate change [74].

## Materials and Methods

**Embryonic experiments.** Animal rearing, P-element mediated transformations, embryonic collections, staging, anti-DigU probe synthesis, and whole-mount *in situ* hybridizations were conducted as previously reported [7].

### Probes for whole-mount *in situ* hybridization in *D. willistoni* embryos.

Primers for probe synthesis are as listed here. *rho*: 5'-CCGCC TTTGC CTATG ACCGT TATAC AATGC and 5'-Pr-TTAGG ACACA CCCAA GTCGT GC, where Pr = the T7 promoter sequence 5'-CCGCC TAATA CGACT CACTA TAGGG. *vn*: 5'-CCGCC TAGTG ACGAC AACAA CAACA GTAGC and 5'-Pr-ATTTT CACTCA CAGCC ATTTT CACC. *vnd*: 5'-CCGCC CTAGT CCGGA TAGCA CTTCG C and 5'-Pr-CGGCT GCCAC ATGTT GATAG G. *brk*: 5'-CCGCC AACAA AGTTC GTCGG CAACA ACG and 5'-Pr-CATGG TGAGG TGAGG ACTAT GG.

**Whole genome sequence analysis.** Current versions for all genomes were downloaded from Flybase ([www.flybase.org](http://www.flybase.org)) and these correspond to assembly versions: dmel ver5.22, dana ver1.3, dpse. ver2.6, dwil ver1.3, and dvir ver1.2. Various whole-genome queries were conducted using shell scripts composed of shell, perl, grep, and wc UNIX commands and are available upon request. Separate queries were conducted for NEE signatures and CA-satellite content. Special genome files were processed for counting percent content of a given motif. We call these “\*.HNF” files because they are header and N-free files; these having been replaced by newline characters.

## Acknowledgments

The authors thank Michael Dietrich, Mark McPeck, Alysha Heimberg, Kevin Peterson, Lisa Fleischer, Ilya Ruvinsky, and Bryan Kolaczowski for reading and commenting on earlier versions of the manuscript. The authors also thank Ann Lavanway for technical assistance. Completion of this work was supported by an NSF CAREER award to A.E. (IOS 0952743).

## References

1. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* 104 Suppl 1: 8605-12.
2. Carroll SB, Prud'homme B, Gompel N (2008) Regulating evolution. *Sci Am* 298: 60-7.
3. Wang X, Chamberlin HM (2002) Multiple regulatory changes contribute to the evolution of the *caenorhabditis* lin-48 ovo gene. *Genes Dev* 16: 2345-9.
4. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* 430: 85-8.
5. Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in *drosophilidae*. *PLoS Biol* 4: e386.
6. McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, et al. (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* 448: 587-590.
7. Crocker J, Tamori Y, Erives A (2008) Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biology* 6: e263.
8. Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, et al. (2008) Human-specific gain of function in a developmental enhancer. *Science* 321: 1346-50.
9. Williams TM, Selegue JE, Werner T, Gompel N, Kopp A, et al. (2008) The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *drosophila*. *Cell* 134: 610-23.
10. Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *drosophila* species. *Nat Genet* 40: 346-50.
11. Shirangi TR, Dufour HD, Williams TM, Carroll SB (2009) Rapid evolution of sex pheromone-producing enzyme expression in *drosophila*. *PLoS Biol* 7: e1000168.
12. Arnone MI, Davidson EH (1997 May) The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124: 1851-1864.
13. Wunderlich Z, Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* 25: 434-40.
14. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *drosophila* genome. *Proc Natl Acad Sci U S A* 99: 757-762.
15. Markstein M, Markstein P, Markstein V, Levine MS (2002) Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the *drosophila* embryo. *Proc Natl Acad Sci U S A* 99: 763-768.
16. Ptashne M (2004) A genetic switch: phage lambda revisited. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, 3rd ed edition.
17. Wolpert L (1989) Positional information revisited. *Development* 107 Suppl: 3-12.
18. Anderson KV, Bokla L, Nusslein-Volhard C (1985) Establishment of dorsal-ventral polarity in the *drosophila* embryo: the induction of polarity by the toll gene product. *Cell* 42: 791-798.
19. Jiang J, Kosman D, Ip YT, Levine M (1991) The dorsal morphogen gradient regulates the mesoderm determinant twist in early *drosophila* embryos. *Genes Dev* 5: 1881-91.

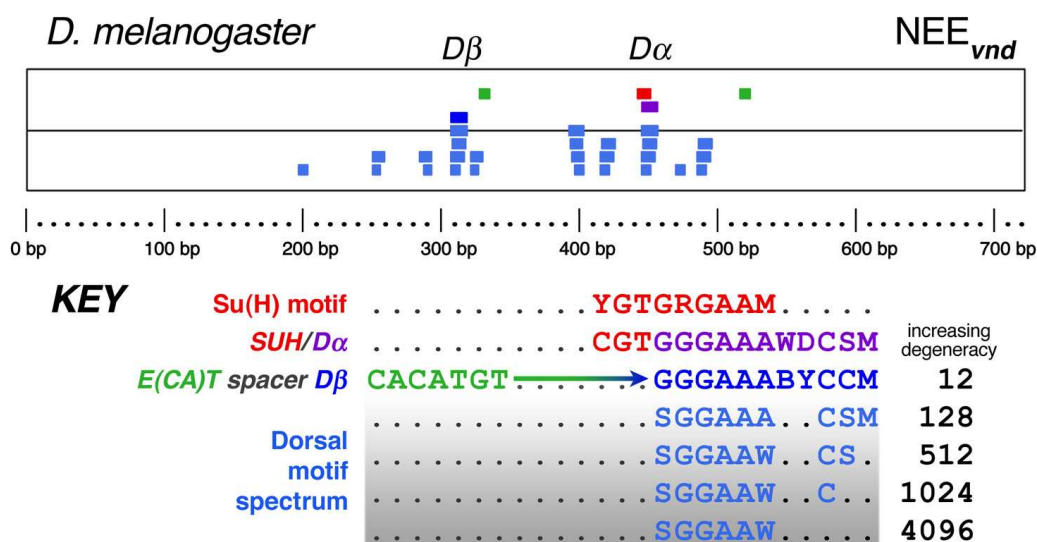
20. Small S, Kraut R, Hoey T, Warrior R, Levine M (1991) Transcriptional regulation of a pair-rule stripe in drosophila. *Genes Dev* 5: 827-39.
21. Ip YT, Levine M, Small SJ (1992) The bicoid and dorsal morphogens use a similar strategy to make stripes in the drosophila embryo. *J Cell Sci Suppl* 16: 33-38.
22. Norris JL, Manley JL (1992) Selective nuclear transport of the drosophila morphogen dorsal can be established by a signaling pathway involving the transmembrane protein toll and protein kinase a. *Genes Dev* 6: 1654-1667.
23. Reinitz J, Mjolsness E, Sharp DH (1995) Model for cooperative control of positional information in drosophila by bicoid and maternal hunchback. *J Exp Zool* 271: 47-56.
24. Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, et al. (2004) Dynamic control of positional information in the early drosophila embryo. *Nature* 430: 368-71.
25. Moussian B, Roth S (2005) Dorsoventral axis formation in the drosophila embryo—shaping and transducing a morphogen gradient. *Curr Biol* 15: R887-99.
26. Gregor T, Tank DW, Wieschaus EF, Bialek W (2007) Probing the limits to positional information. *Cell* 130: 153-64.
27. Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* 130: 141-52.
28. Reinitz J (2007) Developmental biology: a ten per cent solution. *Nature* 448: 420-1.
29. Gregor T, McGregor AP, Wieschaus EF (2008) Shape and function of the bicoid morphogen gradient in dipteran species with different sized embryos. *Dev Biol* 316: 350-8.
30. Papatsenko D, Levine M (2005) Quantitative analysis of binding motifs mediating diverse spatial readouts of the dorsal gradient in the drosophila embryo. *Proc Natl Acad Sci U S A* 102: 4966-71.
31. Zinzen RP, Senger K, Levine M, Papatsenko D (2006) Computational models for neurogenic gene expression in the drosophila embryo. *Curr Biol* 16: 1358-65.
32. Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, et al. (2006) Quantitative and predictive model of transcriptional control of the drosophila melanogaster even skipped gene. *Nat Genet* 38: 1159-65.
33. Brown CD, Johnson DS, Sidow A (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* 317: 1557-1560.
34. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet* 4: e1000106.
35. Liberman LM, Stathopoulos A (2009) Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Dev Biol* 327: 578-589.
36. Cameron RA, Davidson EH (2009) Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Dev Biol* 336: 122-35.
37. Erives A, Levine M (2004 Mar 16) Coordinate enhancers share common organizational features in the drosophila genome. *Proc Natl Acad Sci U S A* 101: 3851-3856.
38. Crocker J, Erives A (2008 Nov) A closer look at the eve stripe 2 enhancers of drosophila and themira. *PLoS Genet* 4: e1000276.
39. Grimaldi DA, Engel MS (2005) Evolution of the insects. Cambridge ; New York: Cambridge University Press. <http://www.loc.gov/catdir/toc/cam051/2004054605.html>.



40. Bertone MA, Courtney GW, Wiegmann BM (2008) Phylogenetics and temporal diversification of the earliest true flies (insecta: Diptera) based on multiple nuclear genes. *Systematic Entomology* 33: 668–687.
41. Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, et al. (2009) Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol* 7: 34.
42. Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M (1992) dorsal-twist interactions establish snail expression in the presumptive mesoderm of the drosophila embryo. *Genes Dev* 6: 1518–1530.
43. Ip YT, Park RE, Kosman D, Bier E, Levine M (1992) The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the drosophila embryo. *Genes Dev* 6: 1728–1739.
44. Jiang J, Levine M (1993) Binding affinities and cooperative interactions with bhlh activators delimit threshold responses to the dorsal gradient morphogen. *Cell* 72: 741–752.
45. Gonzalez-Crespo S, Levine M (1993) Interactions between dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in drosophila. *Genes Dev* 7: 1703–1713.
46. Szymanski P, Levine M (1995) Multiple modes of dorsal-bhlh transcriptional synergy in the drosophila embryo. *EMBO J* 14: 2229–2238.
47. Bailey AM, Posakony JW (1995) Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to notch receptor activity. *Genes Dev* 9: 2609–2622.
48. Ray RP, Schupbach T (1996) Intercellular signaling and the polarization of body axes during drosophila oogenesis. *Genes Dev* 10: 1711–1723.
49. Lecourtois M, Schweisguth F (1997) Role of suppressor of hairless in the delta-activated notch signaling pathway. *Perspect Dev Neurobiol* 4: 305–11.
50. Morel V, Le Borgne R, Schweisguth F (2003) Snail is required for delta endocytosis and notch-dependent activation of single-minded expression. *Dev Genes Evol* 213: 65–72.
51. Gray S, Szymanski P, Levine M (1994) Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev* 8: 1829–38.
52. Cowden J, Levine M (2002) The snail repressor positions notch signaling in the drosophila embryo. *Development* 129: 1785–93.
53. Flores-Saaib RD, Courey AJ (2000) Regulation of dorso/ventral patterning in the drosophila embryo by multiple dorsal-interacting proteins. *Cell Biochem Biophys* 33: 1–17.
54. Bhaskar V, Valentine SA, Courey AJ (2000) A functional interaction between dorsal and components of the smt3 conjugation machinery. *J Biol Chem* 275: 4033–40.
55. Bhaskar V, Courey AJ (2002) The madf-bess domain factor dip3 potentiates synergistic activation by dorsal and twist. *Gene* 299: 173–84.
56. Ratnaparkhi GS, Duong HA, Courey AJ (2008) Dorsal interacting protein 3 potentiates activation by drosophila rel homology domain proteins. *Dev Comp Immunol* 32: 1290–300.
57. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature* 450: 203–218.
58. Castanon I, Von Stetina S, Kass J, Baylies MK (2001) Dimerization partners determine the activity of the twist bhlh protein during drosophila mesoderm development. *Development* 128: 3145–59.

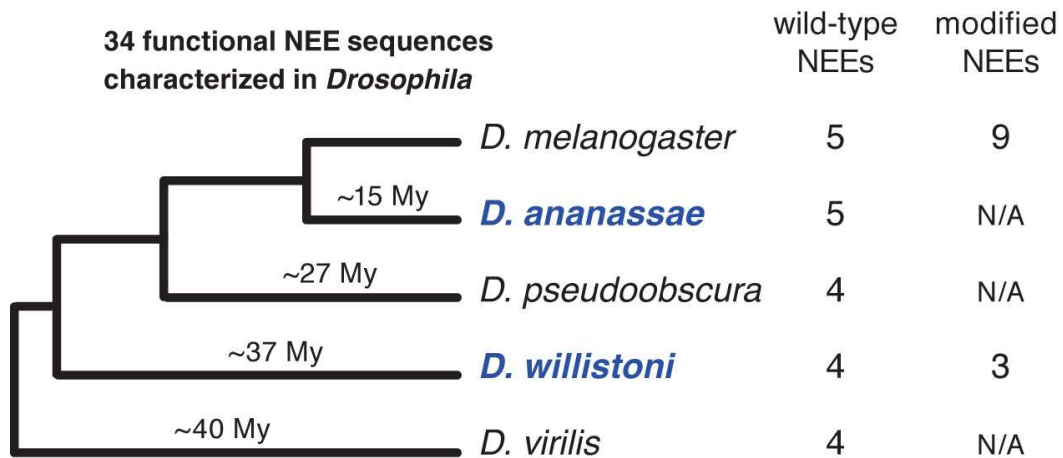
59. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624-6.
60. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96-8.
61. Schlötterer C, Harr B (2000) *Drosophila virilis* has long and highly polymorphic microsatellites. *Mol Biol Evol* 17: 1641-6.
62. Harr B, Zangerl B, Schlötterer C (2000) Removal of microsatellite interruptions by dna replication slippage: phylogenetic evidence from *drosophila*. *Mol Biol Evol* 17: 1001-9.
63. Harr B, Schlötterer C (2000) Long microsatellite alleles in *drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* 155: 1213-20.
64. Lecourtois M, Schweisguth F (1998) Indirect evidence for delta-dependent intracellular processing of notch in *drosophila* embryos. *Curr Biol* 8: 771-4.
65. Peterson BK, Hare EE, Iyer VN, Storage S, Conner L, et al. (2009) Big genomes facilitate the comparative identification of regulatory elements. *PLoS One* 4: e4688.
66. Hong JW, Hendrix DA, Levine MS (2008) Shadow enhancers as a source of evolutionary novelty. *Science* 321: 1314.
67. Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, et al. (2005) The role of binding site cluster strength in bicoid-dependent patterning in *drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 4960-4965.
68. Leung G, Eisen MB (2009) Identifying cis-regulatory sequences by word profile similarity. *PLoS One* 4: e6901.
69. Hewitt G (2003) *Evolution on planet earth: the impact of the physical environment*, Amsterdam: Academic Press, chapter 18: Ice ages, species distributions, and evolution. pp. 339-361.
70. Hewitt GM (2004) The structure of biodiversity - insights from molecular phylogeography. *Front Zool* 1: 4.
71. Hewitt GM (2004) Genetic consequences of climatic oscillations in the quaternary. *Philos Trans R Soc Lond B Biol Sci* 359: 183-95; discussion 195.
72. McPeck MA (2008) The ecological dynamics of clade diversification and community assembly. *Am Nat* 172: E270-84.
73. Overpeck JT, Bartlein PJ, Webb T 3rd (1991) Potential magnitude of future vegetation change in eastern north america: Comparisons with the past. *Science* 254: 692-695.
74. Klausmeyer KR, Shaw MR (2009) Climate change, habitat loss, protected areas and the climate adaptation potential of species in mediterranean ecosystems worldwide. *PLoS One* 4: e6392.

## Figures



**Figure 1. Organization of specialized elements within Dorsal *cis*-spectra of canonical NEEs.**

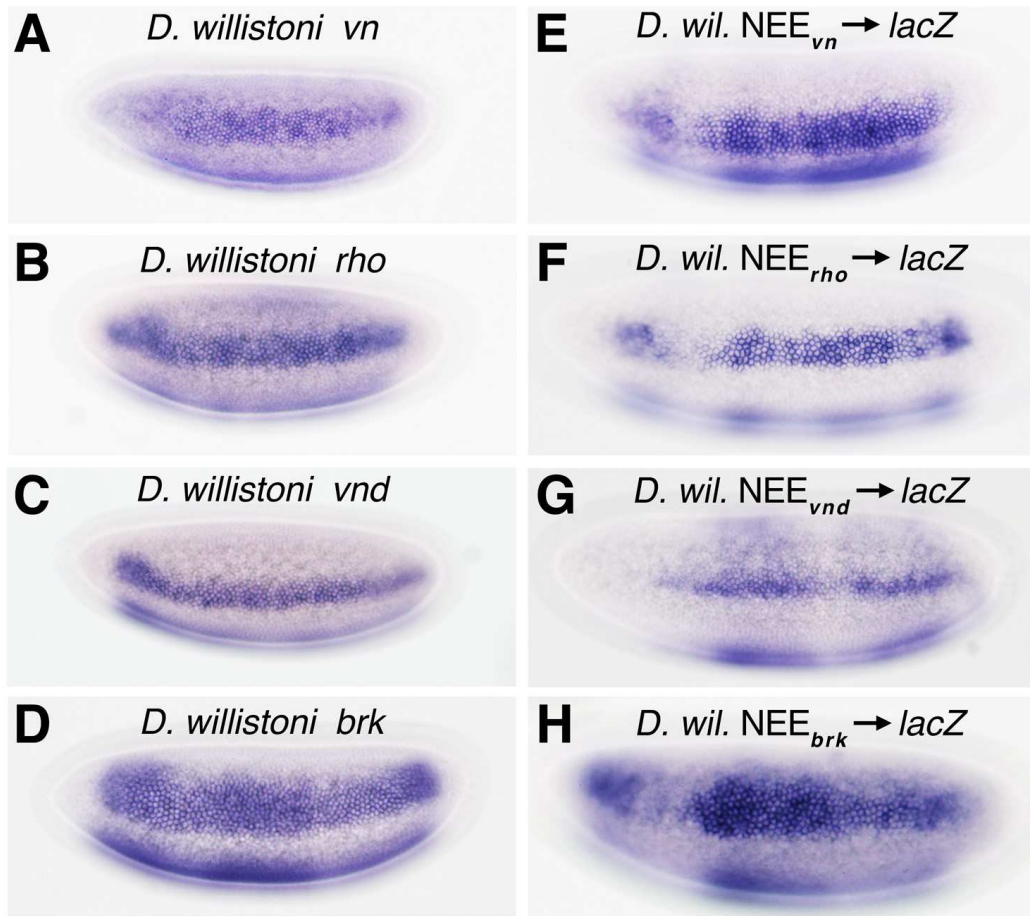
Shown are the specialized sites embedded within the Dorsal *cis*-spectra of the *D. melanogaster* NEE<sub>vnd</sub> sequence, which is representative of canonical NEEs at the *rho*, *vn*, *vnd*, and *brk* loci of the *Drosophila* genus. Numerous lines of evidence in this study demonstrate that the Dorsal *cis*-spectra are specific to mature NEEs (>40 My old), non-functional, and likely produced by dynamic deprecation of precisely spaced Dorsal and Twist sites. Dorsal *cis*-spectra are defined by a motif spectrum of increasingly degenerate Dorsal binding motifs. All instances of the motifs listed in the key are shown in the graphic. The motif sequences in all of the figures and text are written according to IUPAC DNA convention: S = [CG], W = [AT], R = [AG], Y = [CT], K = [GT], M = [CT], B = [CGT], D = [AGT], H = [ACT], V = [ACT], N = [ACGT], where nucleotides in brackets are equivalent. All Dorsal binding sites, motifs, and variants will be depicted with the best halfsite on the 5'- side regardless of its polarity to E(CA)T.



**Figure 2. *Drosophila* phylogeny with tested NEE sequences.**

In this study, we expand our previous studies to two genomes not marked by secondarily-derived compact genome sizes. These genomes correspond to the *D. ananassae* and *D. willistoni* lineages (blue). We also expand our analyses by testing additional mutated versions of these and previously cloned enhancers.





**Figure 3. Functional NEEs from *D. willistoni*.**

Functional NEEs from *D. willistoni* occur in canonical loci that are also expressed in the neurogenic ectoderm. **A–D)** NEE-bearing loci in *D. willistoni* are expressed endogenously in the neuroectoderm of stage 5(2) embryos as shown by *in situ* hybridization with an anti-sense RNA probe to exonic sequences. **E–H)** NEE sequences from *D. willistoni* can drive a *lacZ* reporter gene in transgenic *D. melanogaster* embryos as shown by *in situ* hybridization with an anti-sense RNA probe to *lacZ*. Embryos in all figures are depicted with anterior pole to the left, and dorsal side on top. Image labels indicate the species of the embryo, and the gene or reporter being detected. All reporters are in *D. melanogaster* embryos.

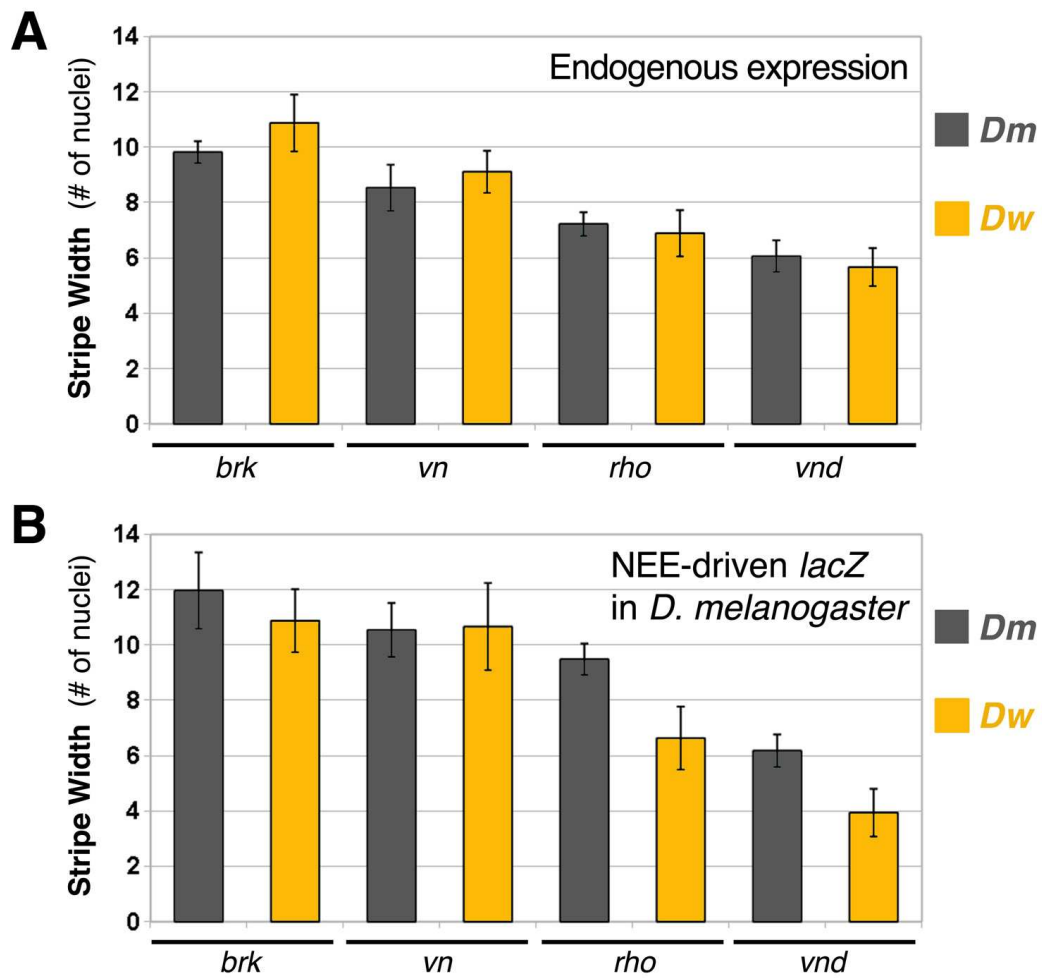
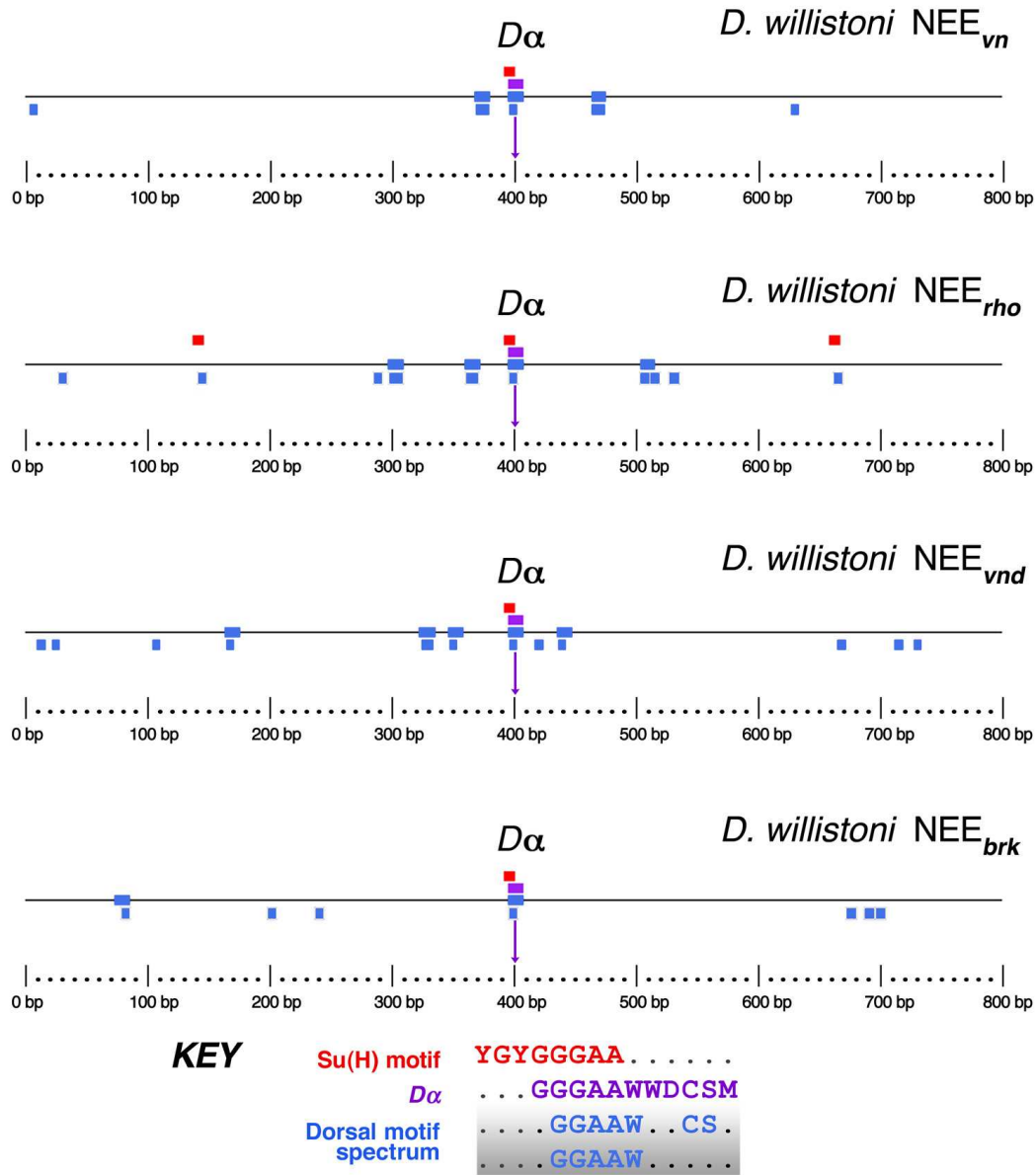
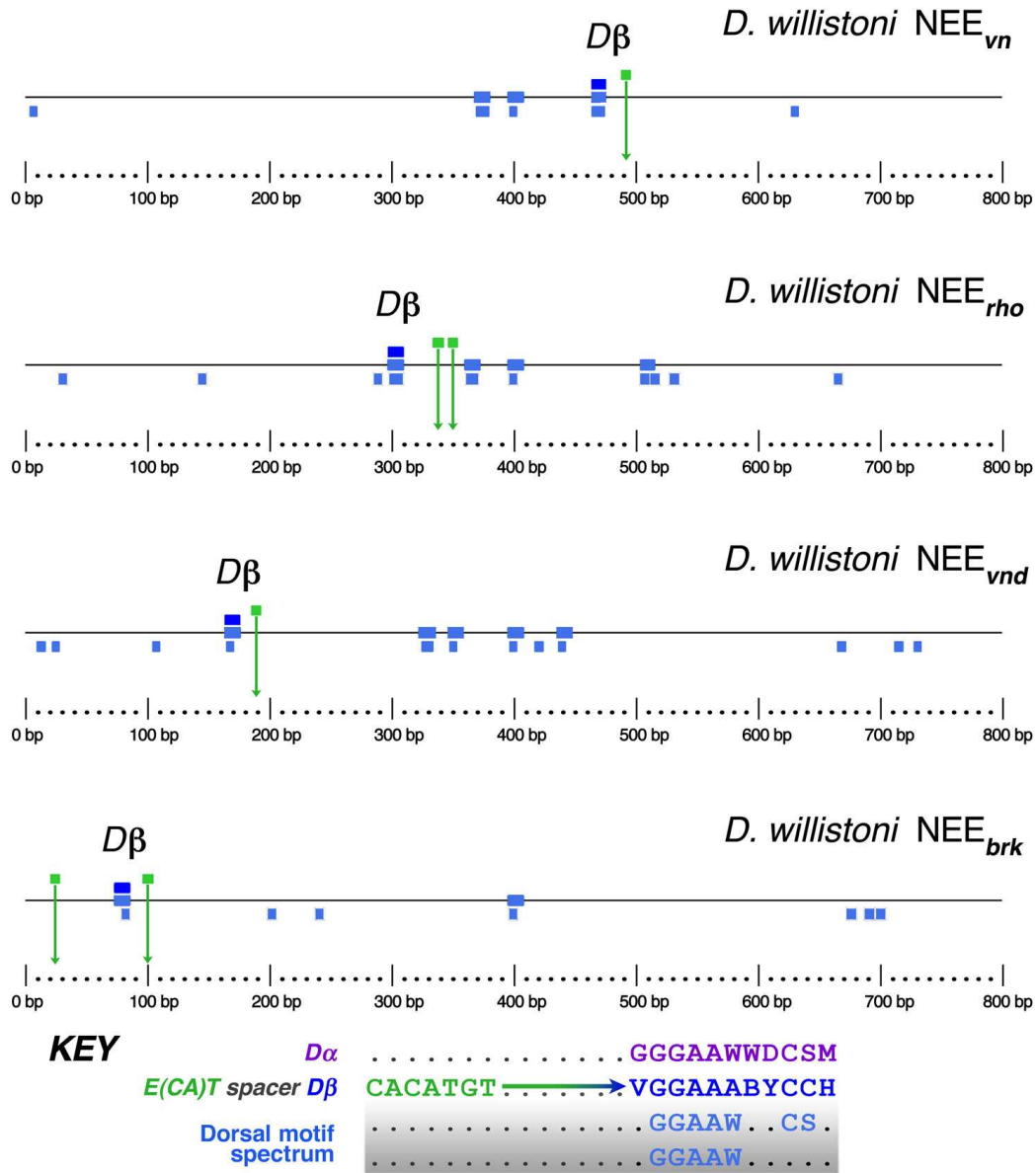


Figure 4. *D. willistoni* NEEs are set to higher concentration thresholds than *D. melanogaster*. See text.



**Figure 5. Dorsal *cis*-spectra in canonical *D. willistoni* NEEs contain a single  $D\alpha$  site.**

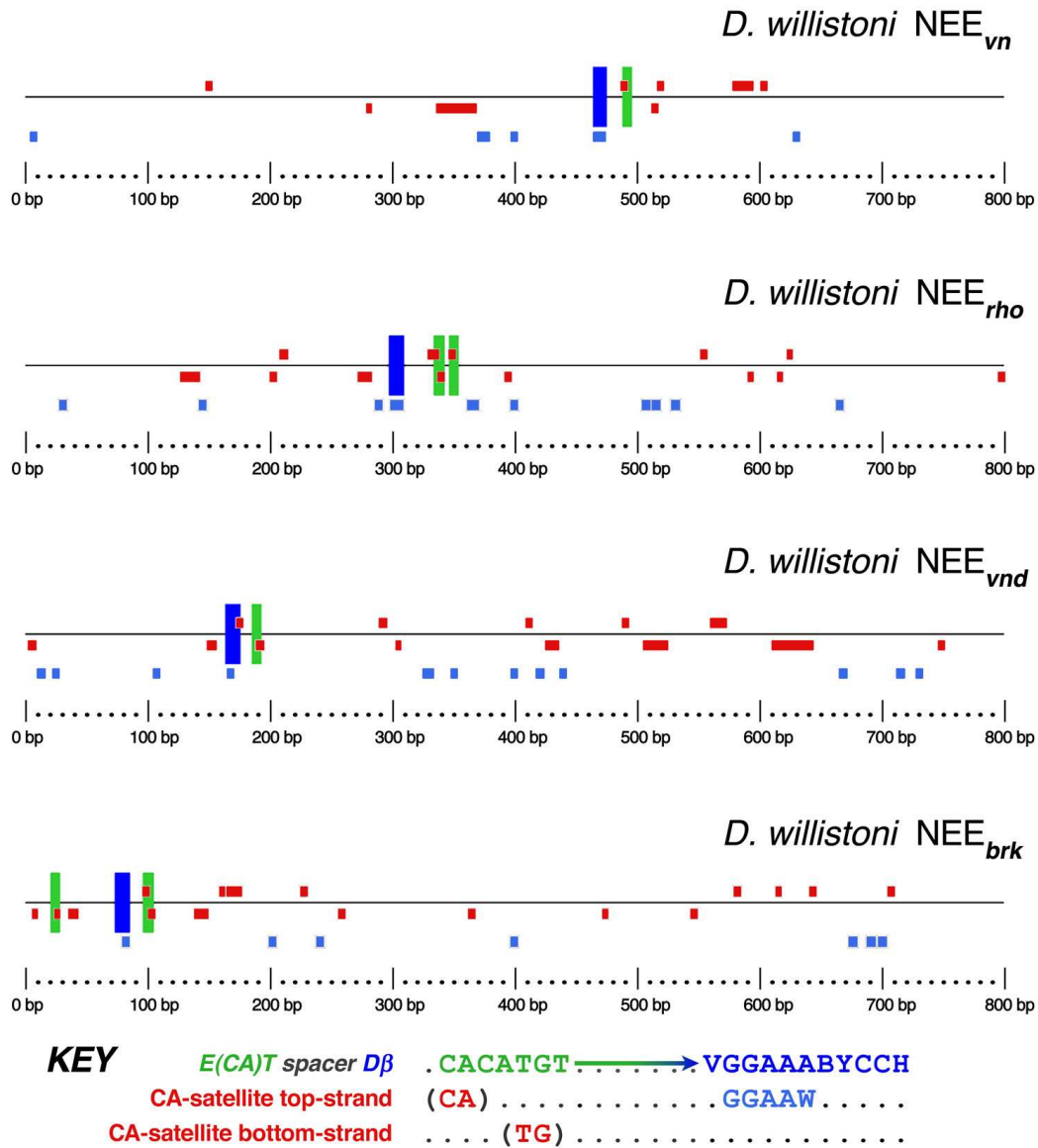
Constituents of Dorsal *cis*-spectra in *D. willistoni* NEEs are visualized by matches to Dorsal halfsites (base *D* halfsite, pale blue) and degenerate full sites (base *D*, light blue) as shown in the key. One such site at each canonical NEE matches the  $D\alpha$  consensus (purple). This same site overlaps a Su(H) binding site (*SUH*, red), which occurs on the top strand at each NEE. For efficient referencing across the set, all NEEs from *D. willistoni* are aligned and centered on the unique  $D\alpha$  site, plus or minus 400 bp, unless otherwise stated.



**Figure 6. Dorsal *cis*-spectra in canonical *D. willistoni* NEEs contain a single *Dβ* site.**

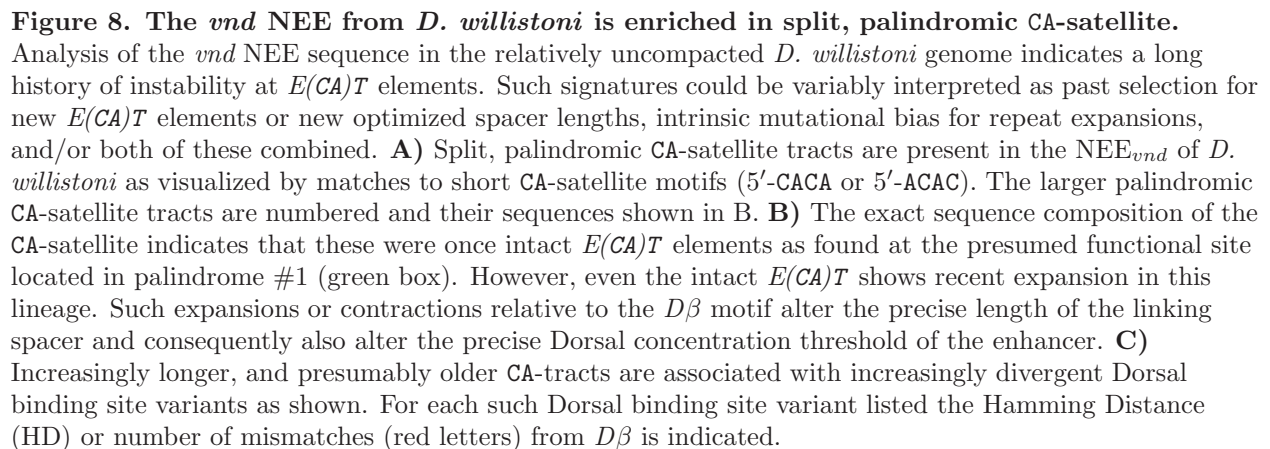
One Dorsal binding site variant in each cluster matches the *Dβ* consensus (dark blue). This specialized *Dβ* site is the closest (<30 bp) Dorsal binding site variant to the *E(CA)T* element (green), which is a binding site for the Dorsal co-activator Twist, and the Snail mesodermal repressor. Sites matching this specialized Dorsal binding motif *Dβ* are distinct from the *Dα* elements (numbered purple labels).

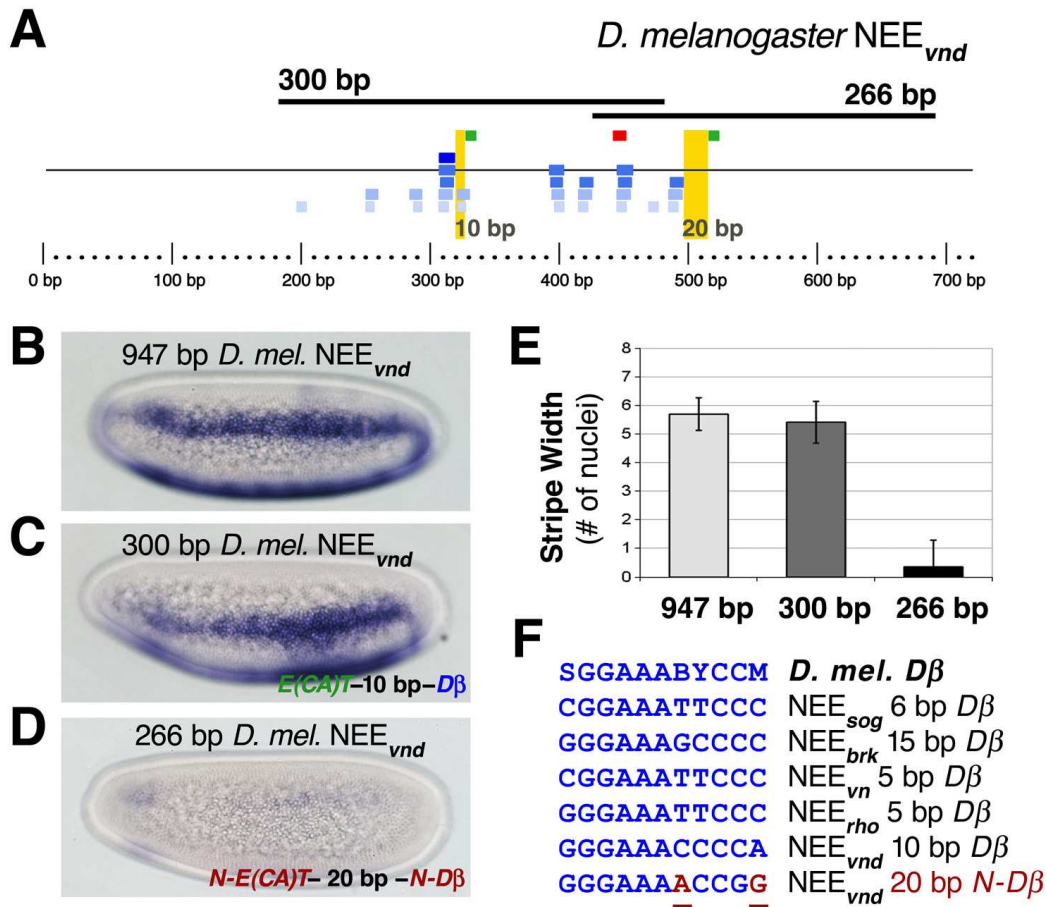




**Figure 7. Canonical NEEs from *D. willistoni* are enriched with CA-satellite.**

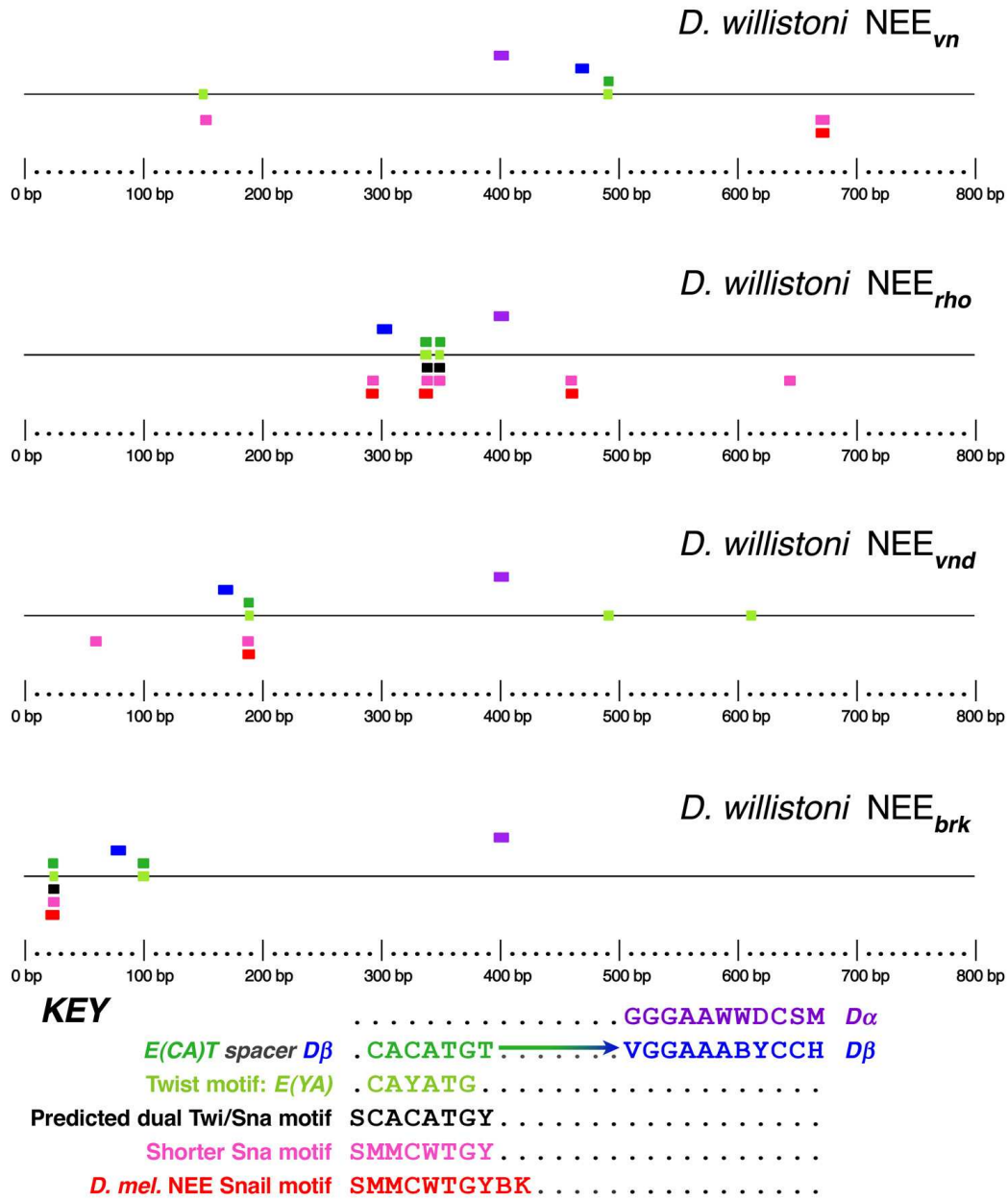
All canonical NEEs from the *D. willistoni* genome also are enriched in CA-satellite, almost as much as the NEE<sub>vnd</sub> sequence, which was present in the latest common ancestor of dipterans (see text). Furthermore, the longest such tracts are associated with divergent *Dβ* halvesites (pale blue). The NEE<sub>vn</sub> *cis*-spectra has expanded CA-satellite tracts associated with ghost *Dβ* elements at ~340–400 bp and again at ~580–630 bp, while NEE<sub>rho</sub> also has expanded CA-satellite tracts coordinated to ghost *Dβ* motifs at ~130–150 bp and again at ~270–290 bp. Such signatures are consistent with the hypothesis that much of the clustering is evidence of past deprecation events between precisely spaced *Dβ* and *E(CA)T* elements. Enhancers are aligned on the unique *SUH/Dα* site at position 400 bp (see Fig. 5).





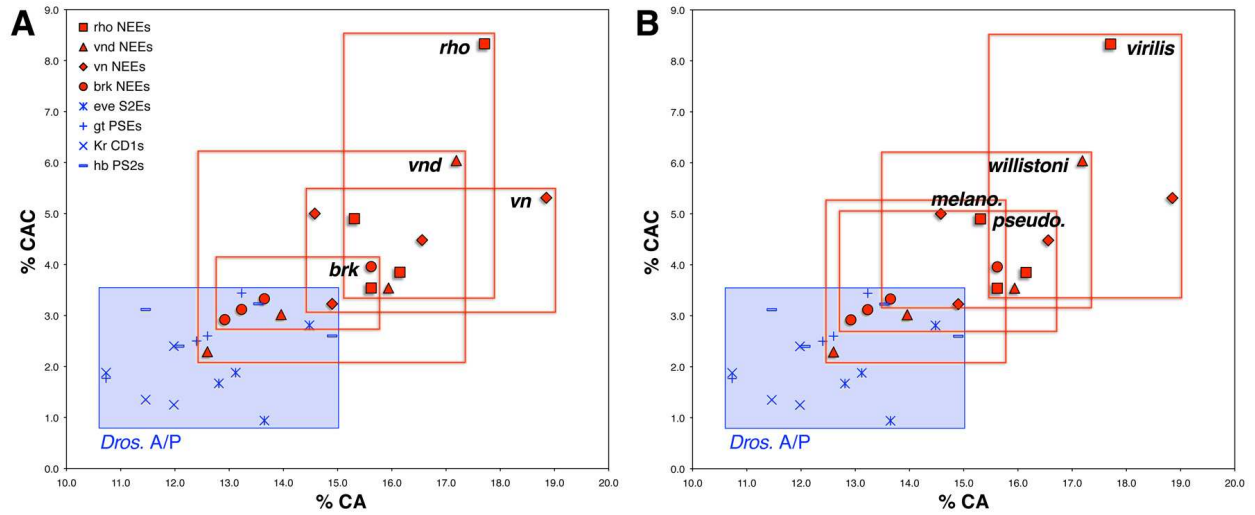
**Figure 9.** The second *E-to-D* encoding within NEE<sub>vnd</sub> was deprecated prior to divergence of the *Drosophila* genus.

**A)** Unlike *D. willistoni*, the NEE<sub>vnd</sub> in *D. melanogaster* has two apparently intact threshold-encodings, one of which is coordinated by a 10 bp spacer (narrow yellow column), and another that is coordinated by a 20 bp spacer (wide yellow column). Motifs follow the key in Fig. 1 except the Dorsal binding spectra are shaded with decreasing intensity as degeneracy increases. The 947 bp “full-length” fragment encompasses the entire 720 bp shown in the graphic. Two smaller tested fragments are shown in dark bold lines. Both of these overlap and include the *SUH/Dα* site (red/blue stack). **B)** The 947 bp NEE<sub>vnd</sub> “full-length” enhancer sequence drives a normal pattern of *lacZ* expression. **C)** The 300 bp NEE<sub>vnd</sub> subfragment drives a similar pattern as the full-length version, despite the absence of the second coordinated Dorsal/Twist binding site pair. **D)** The 266 bp NEE<sub>vnd</sub> subfragment fails to drive a robust lateral stripe of *lacZ* expression at any threshold. Faint staining is occasionally seen in a lateral patch towards anterior pole. **E)** Quantification of the stripe width over several embryos for each construct depicted in A–D shows that the full-length enhancer is not measurably different than the 300 bp fragment containing a single *E-to-D* encoding. **F)** The Dorsal binding site coordinated by 20 bp to the second *E(CA)T* element is divergent (red letters) from the *Dβ* consensus for *D. melanogaster*. This *D. melanogaster* *Dβ* consensus matches the *Dβ* consensi in other lineages more closely than a *D. melanogaster* consensus made with the 20 bp coordinating Dorsal binding site variant.



**Figure 10.** *E(CA)T* versus Twist and Snail binding motifs in *D. willistoni* NEEs.

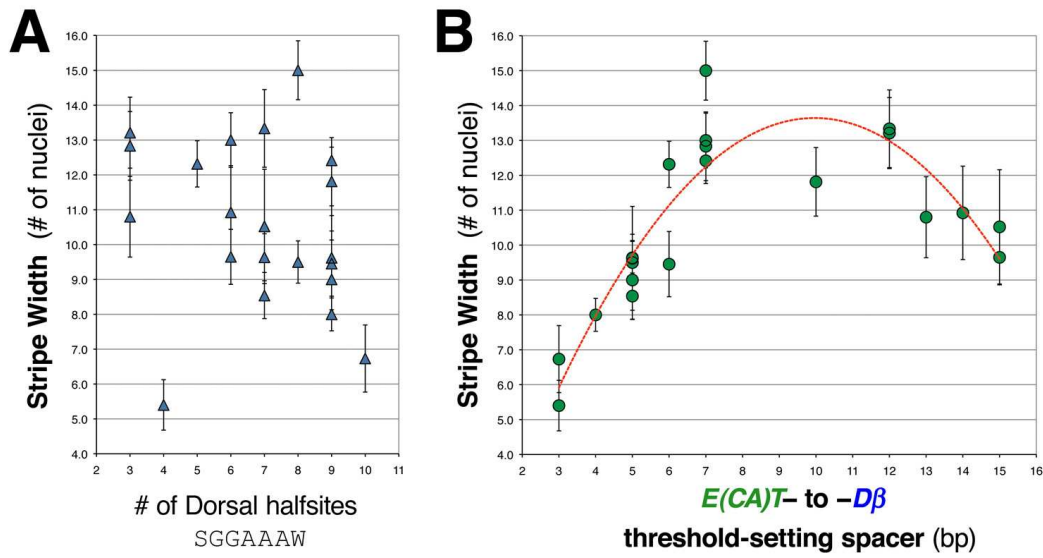
The simple superimposition of motifs representing binding preferences for Twist bHLH complexes and the Snail C<sub>2</sub>H<sub>2</sub> zinc-finger transcriptional repressor, results in a predicted dual motif that is similar but not identical to the observed *E(CA)T* motif. Because the *E(CA)T* motif appears to be subject to repeat expansions and contractions, as seen in Table 2, and because this would result in threshold-modifying variants, we believe that the peculiar difference between the predicted dual site and the observed invariant site, is strong support for our evolutionary model of dynamic deprecation of encodings via CA-satellite instability. These motifs are depicted here.



**Figure 11. High levels of  $E(CA)T$  fragments have accumulated in canonical NEEs across the genus.**

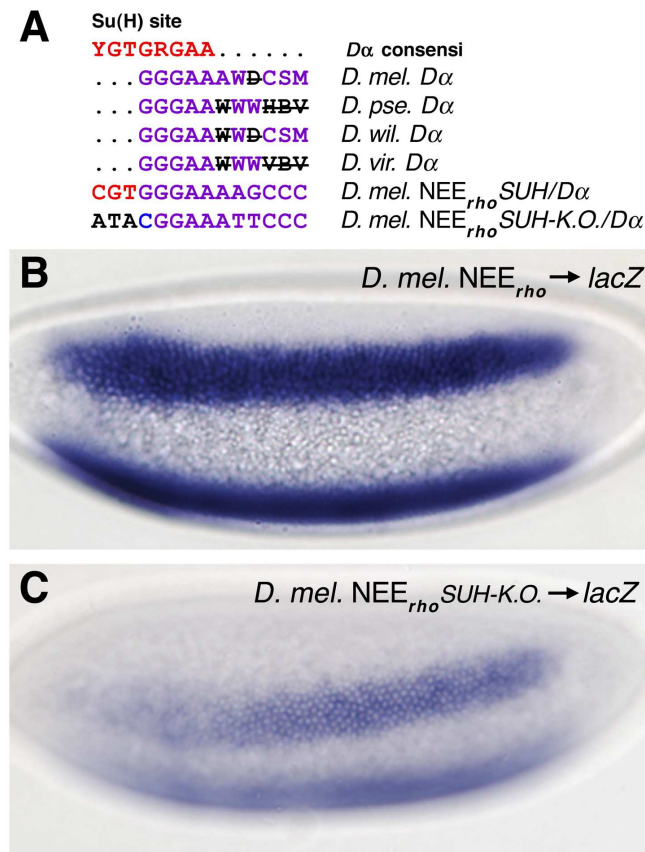
The percentage of sequence that is composed of either 5'-CA dinucleotides or 5'-CAC trinucleotides is graphed for several orthologous groups of enhancers from *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, and *D. virilis*. Each window of NEE sequence is taken  $\pm 480$  bp from *D $\beta$*  for each species. Each window of an A/P enhancer is a 960 bp sequence centered around the Bicoid binding site cluster. **A)** Each orthologous set of NEEs is boxed separately to visualize enrichment relative to other groups. In contrast to the canonical NEEs, the Bicoid binding site clusters of several canonical A/P enhancers at the *eve*, *gt*, *Kr*, and *hb* loci are not associated with high CA-satellite content. All 16 of these enhancers fit within the blue box shown in the graph. **B)** Same as A, except NEEs are boxed by species. Because *D. willistoni* and *D. virilis* represent lineages from each of the subgenera of *Drosophila*, this graph highlights the secondarily-derived, reduced state of CA-satellite in *D. melanogaster* NEEs.



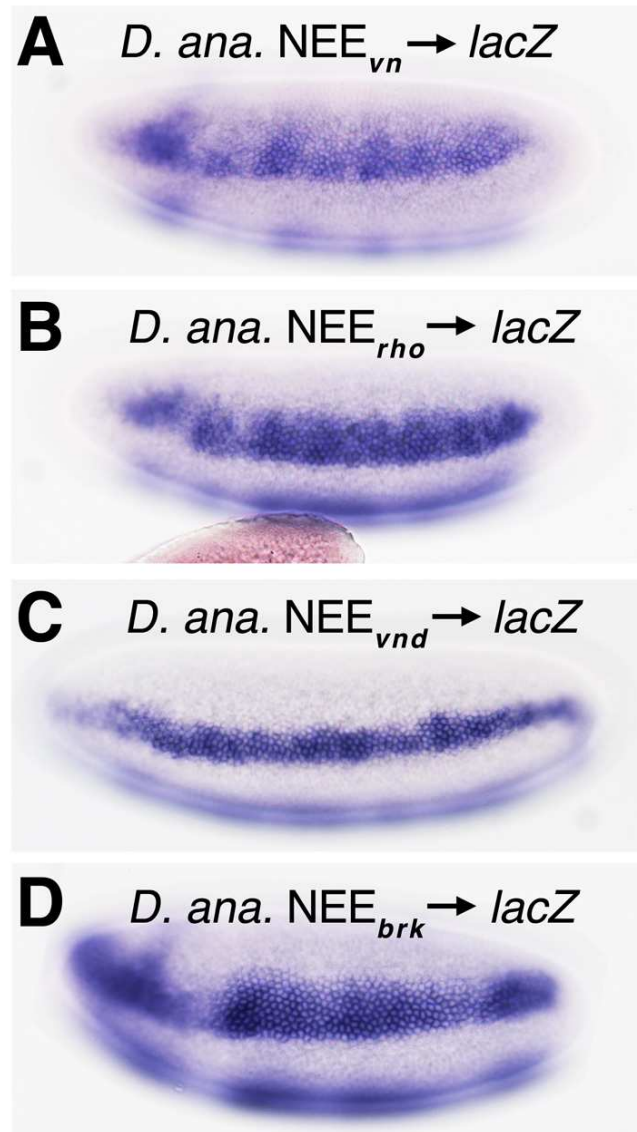


**Figure 12.** The spacer between the *E(CA)T* and *Dβ* encodes the threshold-response to the Dorsal morphogen concentration gradient, and is independent of the number or density of variant Dorsal binding sites.

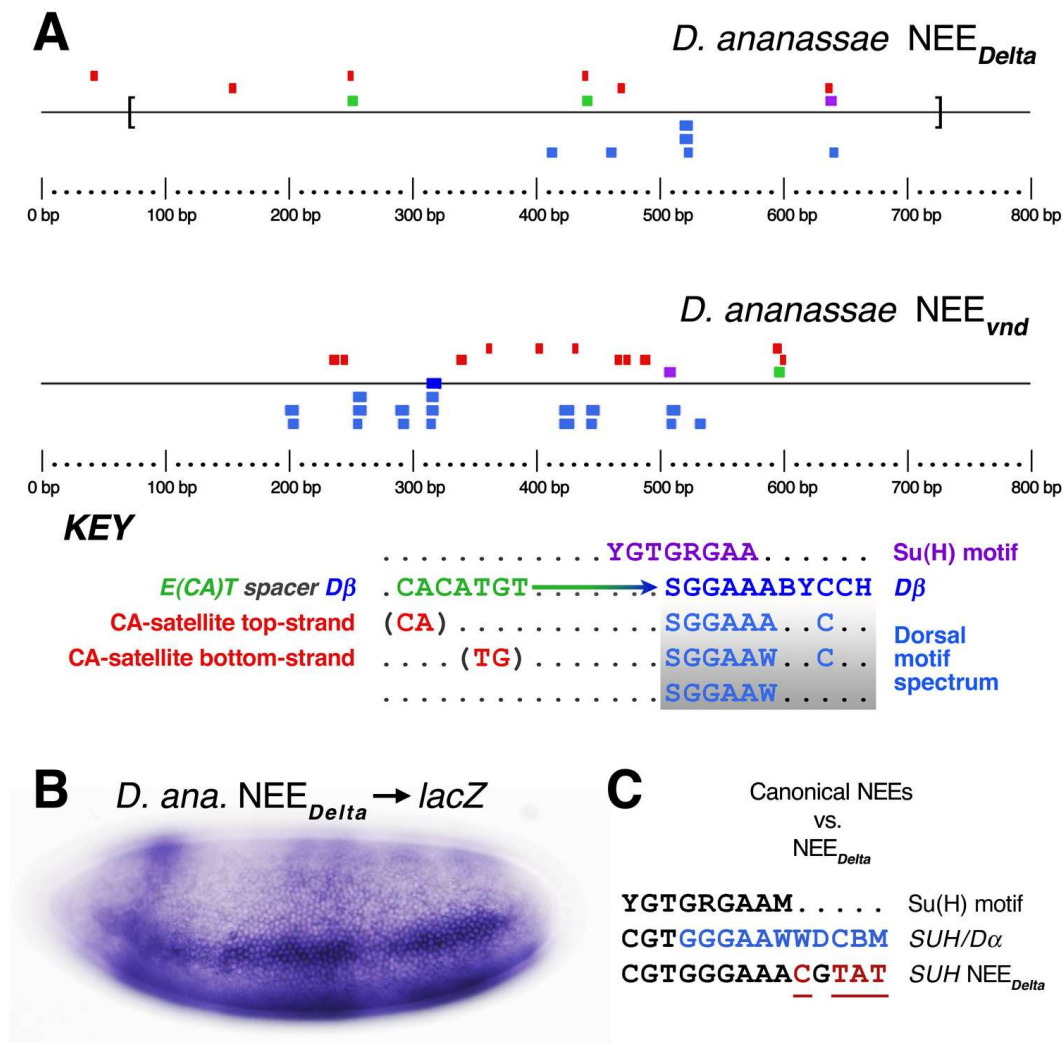
**A)** The number of Dorsal halfsites in the  $\sim 1$  kb window  $\pm 480$  bp from *Dβ* from diverse NEEs of varied age, lineage, and locus, is not predictive of the precise Dorsal concentration threshold readout. **B)** In contrast, the precise spacer length between the *E(CA)T* and *Dβ* elements is predictive (red trendline, second order polynomial) of the precise threshold readout over a range from 3 bp to 15 bp. Vertical axes for both graphs in A and B are aligned for cross-referencing.



**Figure 13.** The *Dα* motif is a *N-Dβ* necro-element that was exapted into a Su(H) binding site. **A)** Alignment of the lineage-specific consensi for *Dα* shows that the portion overlapping the Su(H) binding site portion is the least divergent. The second half of the Dorsal binding site is also increasingly degenerate (black struck-out letters) in comparison to other lineages. Such a signature of divergence is characteristic of drift. Based on this pattern of divergence and the activities of more recent NEEs, we conclude that *Dα* is non-functional and represents a deprecated *Dβ* site exapted into *SUH*. Also shown are the wild-type and mutated sequences of this site tested in the *NEE<sub>rho</sub>* backbone from *D. melanogaster*. **B–C)** Relative activities of *NEE<sub>rho</sub>*-driven reporters differing by the presence (B) or absence (C) of the Su(H) binding site, via a mutation that leaves the Dorsal site intact. The *SUH* element is required for activity levels but not the precise Dorsal concentration threshold encoding. This suggests that Su(H) acts after Dorsal and Twist threshold-activation.

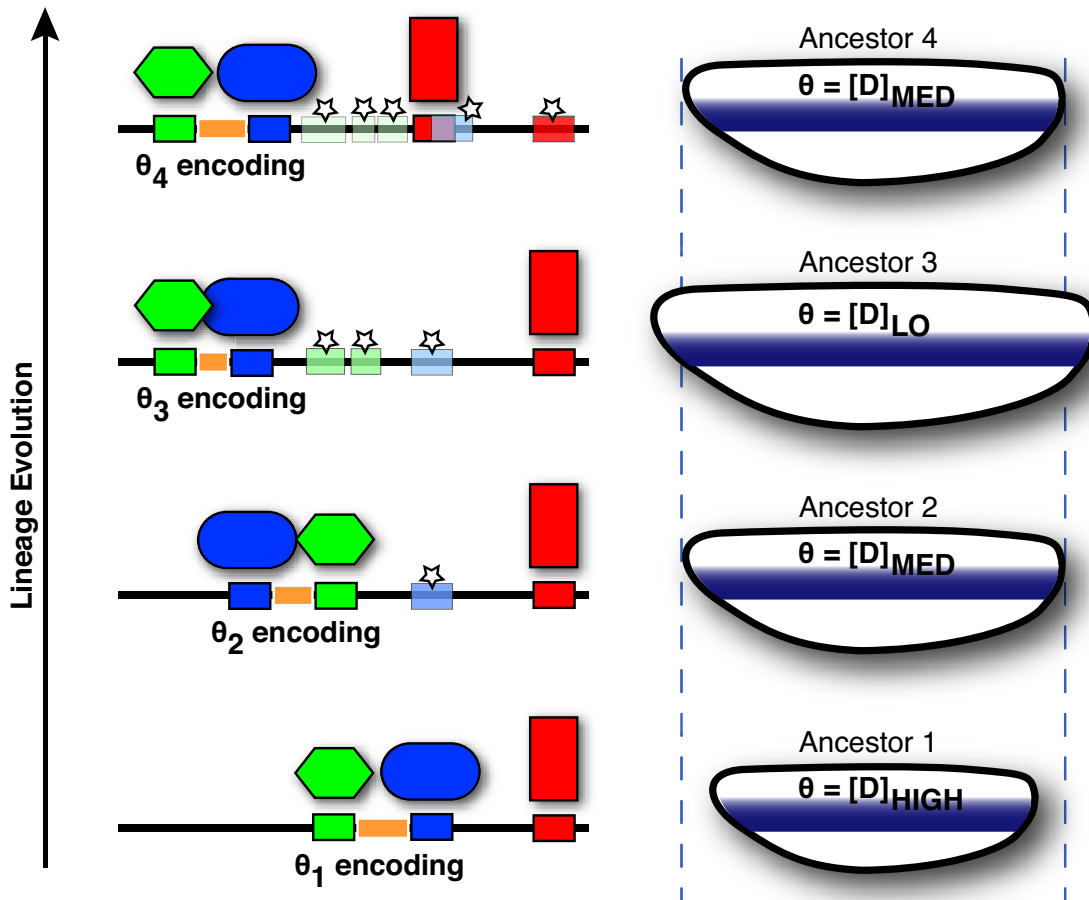


**Figure 14.** Canonical NEEs from *D. ananassae* are functional in *D. melanogaster* embryos. See text.



**Figure 15. A newly evolved NEE at the *Delta* locus of *D. ananassae* has not yet accumulated a necro-element cluster.**

**A)** The genome of *D. ananassae* contains a recently-evolved enhancer NEE<sub>Delta</sub> as well as older, canonical NEEs, such as NEE<sub>vnd</sub> (shown). Dorsal *cis*-spectra are associated with the canonical NEEs but not with the NEE<sub>Delta</sub> sequence, despite employing the essential NEE logic of an *E-to-D* encoding that is near a Su(H) binding site. Brackets in the NEE<sub>Delta</sub> sequence indicate the boundaries of the fragment tested in *D. melanogaster* and shown in B. **B)** The NEE<sub>Delta</sub> module from *D. ananassae* drives a narrow stripe of expression spanning the ~5 nuclei of the mesectoderm and ventral neurogenic ectoderm in *D. melanogaster* embryos. **C)** The *SUH* element does not overlap a ghost *Dα* site. This suggests that the *SUH* element in this recently-evolved NEE sequence is the original site that has not yet needed to re-evolve or track closer to the latest, functioning *E-to-D* encoding. CA-satellite is defined here as sequences matching two CA-dinucleotide repeats or longer (given by the UNIX regular expression: A?(CA){2,}C?).



**Figure 16. Dynamic deprecation produces necro-element clusters over time.**

The evolutionary maintenance of precise threshold encodings via dynamic deprecation and re-selection of replacement encodings can be inferred for Neurogenic Ectodermal Enhancers (NEEs). This process produces necro-element clusters (starred, faded boxes) during the course of lineage evolution. Depicted are binding elements for Dorsal (blue), Twist (green), and Su(H) (red). Spacer elements (orange) separate the Dorsal and Twist elements by a fixed distance, whose length determines the precise threshold encoding required for a given embryo type occurring during lineage evolution. Because genes, such as *vnd*, which are expressed in the neurogenic ectoderm must be expressed over the same number of cells despite evolutionary changes in the size of the embryo (right column), selection will favor NEEs with new, compensatory, threshold encodings (left column). There are multiple other reasons for selecting new threshold encodings, but these are not depicted here for simplicity. New encodings arise either by selection on variant spacer lengths (e.g., evolution of threshold #4), or by the selection of new replacement sites defining preferred spacers (thresholds #1–3). Su(H) sites in particular can also be exapted from relic Dorsal necro-elements when selection favors proximity to the current encoding (see threshold #4). Over time, these processes produce a cluster of necro-elements at an enhancer. Increasingly, this prominent signature heavily influences future evolutionary kinetics.



## Tables

Table 1. Specialized Dorsal motifs in *Drosophila* NEEs.

Species	Motif	Consensus over canonical NEEs
<i>D. melanogaster</i>	<i>SUH</i> / <i>D</i> $\alpha$	CGTGGGAAWDC <u>S</u> M
<i>D. melanogaster</i>	<i>D</i> $\beta$	<u>N</u> VVSGGAABYCC <u>M</u>
<i>D. ananassae</i>	<i>SUH</i> / <i>D</i> $\alpha$	CGTGGGAAWDC <u>B</u> M
<i>D. ananassae</i>	<i>D</i> $\beta$	<u>B</u> SVNGGAABYCCC
<i>D. pseudoobscura</i>	<i>SUH</i> / <i>D</i> $\alpha$	CGTGGGAAW <u>W</u> W <u>H</u> BV
<i>D. pseudoobscura</i>	<i>D</i> $\beta$	<u>B</u> SM <u>S</u> GGAABYCCH
<i>D. willistoni</i>	<i>SUH</i> / <i>D</i> $\alpha$	YGYGGGAAWDC <u>S</u> M
<i>D. willistoni</i>	<i>D</i> $\beta$	<u>D</u> KVSGGAABYC <u>C</u> H
<i>D. virilis</i>	<i>SUH</i> / <i>D</i> $\alpha$	CGTGGGAAW <u>W</u> V <u>B</u> V
<i>D. virilis</i>	<i>D</i> $\beta$	<u>K</u> NVSGGAABYCCH

DNA consensi for the indicated elements of canonical NEEs in each species are listed in IUPAC code. Canonical NEEs are located in *vnd*, *rho*, *vn*, *brk* loci. Underlined letters refer to the more degenerate site of two equivalent positions across the *D* $\alpha$  and *D* $\beta$  consensi for that lineage.

**Table 2.** List of intact or nearly intact encodings in tested NEEs.

No.		Enhancer	$E(CA)T^1$	Spacer	$D\beta^2$
1	878 bp	<i>D. mel.</i> NEE <sub>rho</sub> wt	<u>CACATGT</u>	5 bp	GGGAAATTCCC
2	302 bp	<i>D. mel.</i> NEE <sub>rho</sub> wt min	<u>CACATGT</u>	5 bp	GGGAAATTCCC
3	302 bp	<i>D. mel.</i> NEE <sub>rho</sub> <i>SUH</i> $\Delta$	<u>CACATGT</u>	5 bp	GGGAAATTCCC
4	912 bp	<i>D. mel.</i> NEE <sub>vn</sub> sp -1 bp	<u>CACATGT</u>	4 bp	CGGAAATTCCC
5	913 bp	<i>D. mel.</i> NEE <sub>vn</sub> wt	<u>CACATGT</u>	5 bp	CGGAAATTCCC
6	914 bp	<i>D. mel.</i> NEE <sub>vn</sub> sp +1 bp	<u>CACATGT</u>	6 bp	CGGAAATTCCC
7	915 bp	<i>D. mel.</i> NEE <sub>vn</sub> sp +2 bp	<u>CACATGT</u>	7 bp	CGGAAATTCCC
8	918 bp	<i>D. mel.</i> NEE <sub>vn</sub> sp +5 bp	<u>CACATGT</u>	10 bp	CGGAAATTCCC
9	947 bp	<i>D. mel.</i> NEE <sub>vnd</sub> wt	<u>ACACATGT</u> <u>CACATGTTG</u>	10 bp 20 bp	GGGAAACCCCA GGGAAA $\tilde{A}$ CCG $\tilde{G}$
10	300 bp	<i>D. mel.</i> NEE <sub>vnd</sub> wt trunc	<u>ACACATGT</u>	10 bp	GGGAAACCCCA
11	266 bp	<i>D. mel.</i> NEE <sub>vnd</sub> wt trunc	<u>CACATGTTG</u>	20 bp	GGGAAA $\tilde{A}$ CCG $\tilde{G}$
12	657 bp	<i>D. mel.</i> NEE <sub>brk</sub> wt	<u>CACACATGTGTGTTT</u> <u>CAACACATGTT</u>	15 bp 21 bp	GGGAAAGCCCC GGGAA $\tilde{T}$ GTC $\tilde{A}$ A
13	651 bp	<i>D. mel.</i> NEE <sub>brk</sub> sp -3 bp	<u>CACACATGTGTGTTT</u> <u>CAACACATGTT</u>	12 bp 21 bp	GGGAAAGCCCC GGGAA $\tilde{T}$ GTC $\tilde{A}$ A
14	553 bp	<i>D. mel.</i> NEE <sub>sog</sub> wt	<u>CCACATGTGT</u>	7 bp	CGGAAATTCCC
15	738 bp	<i>D. ana.</i> NEE <sub>rho</sub> wt	<u>CCACATGTGT</u>	3 bp	AGGAAATTCCC
16	758 bp	<i>D. ana.</i> NEE <sub>vn</sub> wt	<u>CACATGT</u>	5 bp	CGGAAATTCCC
17	642 bp	<i>D. ana.</i> NEE <sub>vnd</sub> wt	<u>CACACATGTT</u> <u>CACATGTGTTGG</u>	11 bp 40 bp	GGGAAACCCCC TGGAAA $\tilde{A}$ ACC $\tilde{G}$
18	946 bp	<i>D. ana.</i> NEE <sub>brk</sub> wt	<u>CACACATGTGT</u> <sub>5</sub> GGTTGT	15 bp	TGGAAAGCCCC
19	658 bp	<i>D. ana.</i> NEE <sub>Dl</sub> wt	<u>CACATGTTGCTG</u>	3 bp	GG $\tilde{A}$ AAATTCC $\tilde{A}$
20	843 bp	<i>D. pse.</i> NEE <sub>rho</sub> wt	<u>CACATGTT</u> <u>CCCACATGTGTTT</u> <u>CCCACATGTGTTT</u>	6 bp 19 bp 45 bp	GGGAAATTCCT GGGAAATTCCT CGGAAATTCCT
21	858 bp	<i>D. pse.</i> NEE <sub>vn</sub> wt	<u>CCACATGTTTG</u>	5 bp	CGGAAATTCCC
22	1,305 bp	<i>D. pse.</i> NEE <sub>vnd</sub> wt	<u>CACACATGTTG</u> <u>ACACATGTTTT</u> <u>CACACATGTTGG</u>	11 bp 10 bp 28 bp	GGGAAACTCCA GGGAA $\tilde{T}$ TCCT $\tilde{T}$ GGAAA $\tilde{A}$ ACC $\tilde{G}$
23	859 bp	<i>D. pse.</i> NEE <sub>brk</sub> wt	<u>CACACCACATGTGTGTTT</u>	15 bp	GGGAAAGCCCC
24	784 bp	<i>D. wil.</i> NEE <sub>rho</sub> wt	<u>CACATGT</u> <u>CACACACATGTG</u> <u>CACACACATGTG</u>	6 bp 19 bp 26 bp	GGGAA $\tilde{T}$ TCC $\tilde{T}$ A GGGAA $\tilde{T}$ TCC $\tilde{T}$ A CGGAAATTCCT
25	796 bp	<i>D. wil.</i> NEE <sub>vn</sub> wt	<u>ACAAACACATGT</u>	14 bp	CGGAAATTCCC
26	790 bp	<i>D. wil.</i> NEE <sub>vn</sub> sp -7 bp	<u>CAAAACACATGT</u>	7 bp	CGGAAATTCCC
27	964 bp	<i>D. wil.</i> NEE <sub>vnd</sub> wt	<u>CACACATGTTG</u>	11 bp	GGGAAACCCCA
28	960 bp	<i>D. wil.</i> NEE <sub>vnd</sub> sp + $E(CA)T$	<u>CACATGT</u> <u>CACACATGTTG</u>	7 bp 11 bp	CGGAAA $\tilde{A}$ ACC $\tilde{G}$ GGGAAACCCCA
29	748 bp	<i>D. wil.</i> NEE <sub>brk</sub> wt	<u>CAACACATGTGTTTGGGTG</u>	13 bp	GGGAAAGCCCC
30	742 bp	<i>D. wil.</i> NEE <sub>brk</sub> sp -6 bp	<u>CAACACATGTGTTT</u>	7 bp	GGGAAAGCCCC
31	726 bp	<i>D. vir.</i> NEE <sub>rho</sub> wt	<u>CCACATGTG</u>	7 bp	CGGAAATTCCT
32	828 bp	<i>D. vir.</i> NEE <sub>vn</sub> wt	<u>CCACATGTTTGTG</u>	6 bp	CGGAAATTCCC
33	1,011 bp	<i>D. vir.</i> NEE <sub>vnd</sub> wt	<u>CACACATGTTG</u>	8 bp	GGGAAACCCCA
34	756 bp	<i>D. vir.</i> NEE <sub>brk</sub> wt	<u>CACATGTGTTTGG</u>	12 bp	GGGAAAGCCCC

1. CA-satellite extending from intact  $E(CA)T$  elements is shown when present. Fragmented CA-satellite and their loosely coordinated Dorsal spectra are not shown. Likely deprecated encodings are italicized.

2. Dorsal sites are written with the best halfsite on the top strand.  $D\beta$  sequences departing from species' consensi are indicated with a tilde.

**Table 3.** CA-satellite content in *Drosophila* genomes and their canonical NEE sets.

	<i>D. melanogaster</i> release 5.22	<i>D. willistoni</i> release 1.3	<i>D. virilis</i> release 1.2
Total DNA in assembly	162,370,174 bp	223,610,028 bp	189,205,863 bp
% CA-satellite - genome	3.9%	4.0%	4.5%
% CA-satellite - canonical NEEs	5.3%	7.7%	10.0%
% <i>E(CA)T</i> - canonical NEEs	1.3%	1.5%	1.8%

CA-satellite was defined as CA-dinucleotide repeats of 2 or more with an optional single nucleotide extension of the repeat pattern at either end. Canonical NEE sequences for *vnd*, *rho*, *vn*, *brk* loci were extracted  $\pm 480$  bp from *D $\beta$* .